REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
28-08-2007	DISSERTATION	

4. TITLE AND SUBTITLE
MODELS FOR SERIALLY CORRELATED, OVER OR UNDERDISPERSED, UNEQUALLY SPACED LONGITUDINAL COUNT DATA WITH APPLICATIONS TO ASTHMA INHALER USE

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
MAJ BRUCE STEPHANIE L

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF COLORADO HSC

8. PERFORMING ORGANIZATION REPORT NUMBER
CI07-0065

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
THE DEPARTMENT OF THE AIR FORCE
AFIT/ENEL, BLDG 16
2275 D STREET
WPAFB OH 45433

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Unlimited distribution
In Accordance With AFI 35-205/AFIT Sup 1

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		78	19b. TELEPHONE NUMBER *(Include area code)*

Bruce, Stephanie L. (Ph.D., Analytic Health Sciences, Biostatistics)

Models for Serially Correlated, Over or Underdispersed, Unequally Spaced Longitudinal Count Data with Applications to Asthma Inhaler Use

Thesis directed by Associate Professor Gary K. Grunwald.

This research focuses on longitudinal count data methods that do not conform to the well-behaved properties of normality. Potential complications that can arise with longitudinal count data are serial correlation, subject heterogeneity, underdispersion (or overdispersion), and unequally spaced or missing data. Many of the current models in the literature address one or two of the potential complications, but currently there is not a model that addresses all of the complications listed previously, so our goal, given enough data, was to develop a model capable of accommodating all the listed complications. We applied this model to a National Jewish Medical and Research Center study of asthma inhaler use in asthmatic children during school. Using a likelihood based approach, the data were clearly underdispersed relative to the Poisson distribution so a generalized Poisson process was used. Physical activity was the most influential independent variable resulting in a decrease of inhaler usage when the children did not participate in gym class.

The form and content of this abstract are approved. I recommend its publication.

Approved: Gary K. Grunwald

Bruce, Stephanie L. (Ph.D., Analytic Health Sciences, Biostatistics)

Models for Serially Correlated, Over or Underdispersed, Unequally Spaced Longitudinal Count Data with Applications to Asthma Inhaler Use

Thesis directed by Associate Professor Gary K. Grunwald.

This research focuses on longitudinal count data methods that do not conform to the well-behaved properties of normality. Potential complications that can arise with longitudinal count data are serial correlation, subject heterogeneity, underdispersion (or overdispersion), and unequally spaced or missing data. Many of the current models in the literature address one or two of the potential complications, but currently there is not a model that addresses all of the complications listed previously, so our goal, given enough data, was to develop a model capable of accommodating all the listed complications. We applied this model to a National Jewish Medical and Research Center study of asthma inhaler use in asthmatic children during school. Using a likelihood based approach, the data were clearly underdispersed relative to the Poisson distribution so a generalized Poisson process was used. Physical activity was the most influential independent variable resulting in a decrease of inhaler usage when the children did not participate in gym class.

The form and content of this abstract are approved. I recommend its publication.

Approved: Gary K. Grunwald

MODELS FOR SERIALLY CORRELATED, OVER OR UNDERDISPERSED, UNEQUALLY SPACED LONGITUDINAL COUNT DATA WITH APPLICATIONS TO ASTHMA INHALER USE

By

STEPHANIE L. BRUCE

B.S., Clarkson University, 1995

M.Stat., North Carolina State University, 2001

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Preventive Medicine and Biometrics

2007

The views expressed in this article are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

This thesis for the Doctor of Philosophy degree by

Stephanie L. Bruce

has been approved for the

Department of Preventive Medicine and Biometrics

by

Richard H. Jones, Chair

Gary K. Grunwald, Advisor

Nathan Rabinovitch

Matthew Strand

Stanley Xu

Date 8/15/2007

ACKNOWLEDGEMENTS

As I approach the conclusion of my graduate studies, there are many people I would like to thank. First, I would like to thank my advisor, Gary K. Grunwald, for all his time and effort in helping me achieve my PhD in three years. I would not have finished on time without his patience and guidance. I would also like to acknowledge my committee members: Richard H. Jones, Nathan Rabinovitch, Matthew Strand, and Stanley Xu, who were always encouraging and willing to help whenever I asked. In addition, I am grateful to Gary O. Zerbe for his time and participation on my committee. Overall, I truly enjoyed my courses at UCHSC and appreciate the professors who taught them.

I would also like to thank the United States Air Force Academy Department of Mathematical Sciences for their sponsorship and confidence in my abilities to complete the PhD.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

INTRODUCTION

1.1. Background

The topic of this dissertation is statistical methods for analyzing longitudinal count data. Count data is a specific case of discrete data where the occurrences of events are counted so that the counts are non-negative integers. Common examples of count data include the number of epileptic seizures experienced in one week or the number of times an asthma inhaler is used in one day. When the event is counted over a specified period of time repeatedly on a number of subjects, such as weekly seizure counts for two months or daily inhaler usage counts for one month, this is referred to as longitudinal count data or sometimes in the econometric literature as panel count data.

Longitudinal count data occur frequently in biomedical studies and present challenges in analysis because they do not conform to the well-behaved properties of normality. Potential complications that can arise with longitudinal count data are serial correlation, subject heterogeneity, underdispersion (or overdispersion) relative to the Poisson distribution, and unequally spaced or missing data. There are many existing methods for analyzing longitudinal count data including Liang and Zeger (1986), Lambert (1996), Jorgensen et al (1999), Henderson and Shimakura (2003), and Xu, Jones, Grunwald (2007). These methods are reviewed in Section 3.1. Many of the current models in the literature address one or two of the potential complications, but currently there is not a model that addresses all of the complications listed above. Additionally, it is unclear which models are most appropriate for various longitudinal count data situations because diagnostic tools for these models are not well developed.

This dissertation was motivated by a study of asthma inhaler usage by students enrolled at The Kunsberg School on the National Jewish Medical and Research Center (NJMRC) site. The Kunsberg School, located in Denver, Colorado, is a unique day school program for children who may require medical assistance during the school day. A detailed description of the study is given in the next section.

1.2. Example: Daily asthma inhaler use in grade school children

A NJMRC study of children with asthma measured the number of times a day each subject used their inhaler to relieve asthma symptoms. Children in the study were provided with an inhaler that contained an electronic counter that recorded the number of usages in a 24-hour time period. We used only the inhaler counts made during the school day (approximately 8 A.M. to 2:45 P.M.) when the student attended classes. Measurements were collected for several consecutive months during the school year. Students in the study ranged in age from six to thirteen years old and each student had previously been diagnosed with asthma by a physician. An example of a single subject's data is shown in Figure 1.

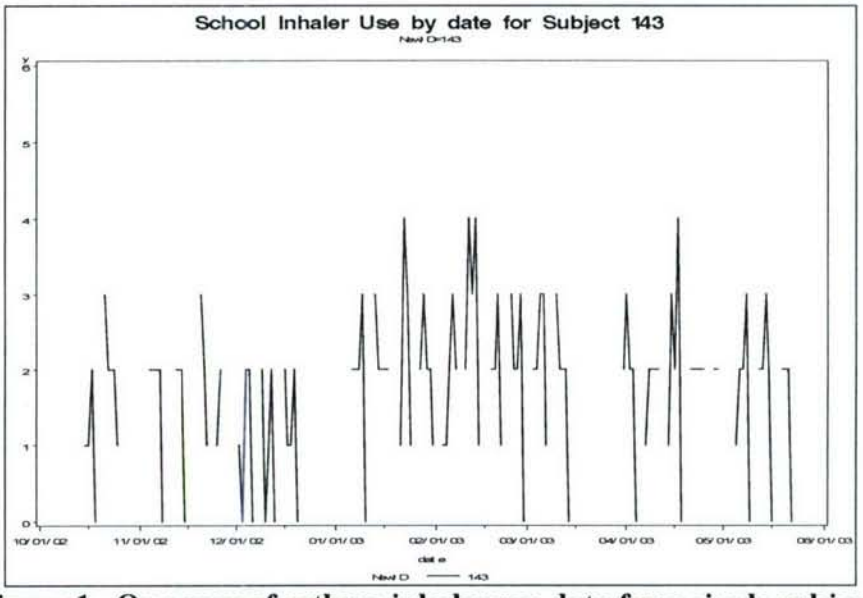


Figure 1. One year of asthma inhaler use data for a single subject.

Inhaler usage is a discrete measure of counts since each usage is counted as an event and the total number of uses for each day of attended school is recorded automatically by the inhaler. In actuality, each inhaler usage is theoretically supposed to be two puffs and odd numbers of puffs should not occur, but occasionally they did. Therefore, the data were transformed into inhaler uses so that a daily count of one or two puffs counted as one use, three or four puffs equaled two inhalers uses, five or six puffs totaled three uses, and so on. The numbers of inhaler uses were recorded each day for numerous subjects over time, resulting in this being longitudinal count data. Several factors can potentially contribute to asthma symptoms worsening, including daily weather variables such as barometric pressure, temperature, humidity, air pollution and especially particulate matter in the air, as well as physical activity and illness (e.g. a cold).

For these data, summary statistics were calculated for each subject's inhaler use counts and the expected values and variances by subject indicate the data are not Poisson distributed. Specifically, many subjects appear to have underdispersed data relative to the Poisson distribution, while a small number of subjects seem to have overdispersed data relative to the Poisson distribution. Figure 2 below plots subject means on the horizontal axis versus subject variances on the vertical axis. The diagonal line represents equality between a subject's mean and variance, while subjects with a variance smaller than their mean fall below the diagonal and are considered underdispersed. The larger group of subjects that are clearly underdispersed and have means that primarily range between one and two inhaler uses per day, is in many cases those who pretreat their asthma before gym class. The majority of subjects are recommended to use their inhalers

prior to physical activity, while a handful of subjects likely do not follow this treatment process due to other medications or less severe asthma.

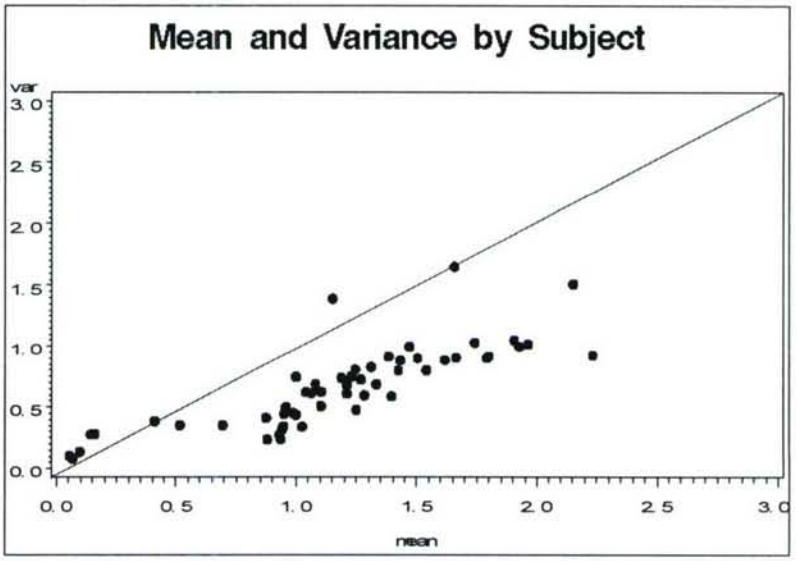


Figure 2. Mean versus variance of number of daily asthma inhaler uses during the school day by subject for 54 children at NJMRC.

We chose to focus on the underdispersed subjects since these data are not easily handled by existing methods. The availability of up to 125 days of data on 54 subjects provides sufficient data to model this effect in addition to serial correlation and possible subject heterogeneity if it exists. All children have some missing days of measurement due to weekends, holidays, or absences from school. The dataset appears to contain many of the complications with longitudinal count data which are discussed in detail in the subsequent section.

1.3. Potential complications with longitudinal count data

Data collected over time on multiple subjects or objects allow investigators to learn additional information about their study beyond a single snapshot in time, but if the analysis of the data does not take into account the longitudinal features within the data the additional information can be lost or misrepresented. Common longitudinal data challenges regularly seen in both Guassian data and count data are serial correlation,

subject heterogeneity, and missing or unequally-spaced data. When the longitudinal data outcome is a count, the additional complication of over- or underdispersion may arise since the mean and variance are typically connected by a common parameter. For example, in the inhaler use dataset subjects tended to use there inhaler consistently resulting in a little variation so that the variance was smaller than the mean for most subjects. These challenges are discussed further in the following subsections.

1.3.1. Serial correlation

For data collected over time on the same subject, serial correlation is often the primary complication. Serial correlation occurs when measurements taken closer together in time have a stronger relationship than measurements taken further apart in time. Subsequently, measurements values tend to be followed by similar values. Serial correlation can be incorporated into a model in various ways, for example by using a latent variable (e.g. an AR (1) process) or by conditioning (i.e. regressing) on the last observed count.

1.3.2. Subject heterogeneity

When a series of measurements collected over time on one subject tend to be high (or tend to be low) in value relative to a series of the same measurements on another subject or the population mean, then subject heterogeneity exists; this is often referred to as between subject variation. It may not always be possible to include both serial correlation and subject heterogeneity in models due to small sample sizes, and in some instances subject heterogeneity can also explain the serial correlation (Jones, 1990, for a normally-distributed outcome).

1.3.3. Over- or underdispersion

Count data are almost always associated with the Poisson distribution, which has the unique property that the expected value and variance are equal. Overdispersion relative to the Poisson distribution occurs when the variance exceeds the mean by an amount beyond what is expected due to random error and underdispersion is when the count variance is less than the mean. This dissertation focuses on underdispersion relative to the Poisson distribution since it is less well studied. Also, underdispersion occurs in the dataset that motivated this work (the asthma inhaler counts), but the model proposed in Section IV is also capable of handling excess variability as well as Poisson data.

1.3.4. Missing or unequally spaced data

Missing data can occur for assorted reasons, including equipment malfunction, human error, or practical difficulties such as subject unavailability or days that are simply planned not to have data collection. Missing longitudinal count data create gaps in the observed information which results in unequally spaced data. Often unequally spaced data assume that measurement intervals are of equal length and some observations are missing (e.g. a missing day of observations), but measurement intervals may be of different lengths as well (e.g. some counts accumulated over two days versus other counts that only represent one day).

In this dissertation I propose a new model that is dynamic enough to handle all the potential longitudinal count data complications described above. This model modifies an existing serially correlated count data model by accommodating subject heterogeneity and underdispersion (or overdispersion) in a state-space model which uses a modified

Kalman recursion process for estimation. The model is applied to the NJMRC asthma inhaler use data, and methods for assessing model appropriateness are presented.

CHAPTER II

COUNT DATA MODELS

There is a vast amount of literature that discusses a range of methods and distributions for analyzing count data depending on the covariates, the presence of heterogeneity or correlation, if the outcome is collected only once or repeatedly over time, and if the equidispersion assumption of the Poisson distribution is viable. This chapter will focus on methods for cross-sectional count data with emphasis on over- and underdispersion.

2.1. The Poisson distribution

The standard distribution for count data is the Poisson distribution, which was originally derived by Siméon-Denis Poisson, a French mathematician in 1837. Since the Poisson distribution is the point of reference for all count data models, a review of this distribution is given here. The probability for y occurrences of an event is:

$$\mathrm{P}(Y = y \mid \mu) = \frac{e^{-\mu}\mu^{y}}{y!}; \qquad y = 0,1,2,\ldots;\ \mu > 0$$

where μ is the expected number of occurrences, $\mathrm{E}(Y) = \mu$, as well as the variance, $\mathrm{V}(Y) = \mu$. μ is also referred to as the rate of events per unit time. Therefore, for Poisson distributed data, as the mean grows the variance grows at an equal rate; it is often this property of the distribution that is violated for count data. The next two sections discuss methods for count data when the assumption of equal mean and variance is inaccurate.

2.2. Overdispersed count data models

The common approach to model overdispersion for count data is to allow the mean of the Poisson distribution to be random. Typically, this is done by allowing the mean to follow a gamma or lognormal distribution.

2.2.1. The Poisson-gamma mixture model

A Poisson-gamma mixture model conditions the counts on a gamma distributed mean and the conditional counts are Poisson distributed; the marginal distribution of the counts is then a negative binomial distribution. The hierarchy of the model is (Casella and Berger, 2002):

$$Y \mid \mu \sim \text{Poisson}(\mu)$$

$$\mu \sim \text{Gamma}(\alpha, \beta)$$

$$Y \sim \text{Negative Binomial}\left(\alpha, \frac{1}{(1+\beta)}\right).$$

This mixture model results in a closed form solution with marginal (or unconditional) mean,

$$\text{E}(Y) = \frac{\alpha\left(1 - \frac{1}{(1+\beta)}\right)}{\frac{1}{(1+\beta)}} = \frac{\alpha\left(\frac{\beta}{(\beta+1)}\right)}{\frac{1}{(\beta+1)}} = \alpha\beta,$$

and marginal variance,

$$\text{V}(Y) = \frac{\alpha\left(1 - \frac{1}{(1+\beta)}\right)}{\left(\frac{1}{(1+\beta)}\right)^2} = \frac{\alpha\left(\frac{\beta}{(\beta+1)}\right)}{\left(\frac{1}{(\beta+1)}\right)^2} = \alpha\beta(\beta+1) = \alpha\beta + \alpha\beta^2,$$

which are derived directly from the negative binomial distribution. The unconditional mean and variance of Y can also be found using the rules of conditional expectation and variance (e.g. Casella and Berger, 2002, p. 164 and 167 respectively):

$$E(Y) = E(E(Y \mid \mu)) = E(\mu) = \alpha\beta$$

and

$$V(Y) = E(V(Y \mid \mu)) + V(E(Y \mid \mu)) = E(\mu) + V(\mu) = \alpha\beta + \alpha\beta^2 .$$

The latter equation shows that the variance is made up of a first piece due to the Poisson variation and a second piece due to the variation in the Poisson mean, which leads to a variance larger than Poisson variance or overdispersion.

2.2.2. The Poisson-lognormal mixture model

Another choice to model overdispersion for counts is to use a lognormal distributed mean in place of the gamma distributed mean described in Section 2.2.1. This does not result in a closed form solution for the marginal distribution, but is easily handled with modern computing methods.

$$Y \mid \mu \sim \text{Poisson}(\mu)$$

$$\ln(\mu) \sim \text{Normal}(\zeta, \sigma^2)$$

The mean of the conditional distribution of Y is specified to have a lognormal distribution as described above, and μ can be shown to have mean $E(\mu) = \exp\left(\zeta + \frac{1}{2}\sigma^2\right)$. This leads to an unconditional mean of (Casella and Berger, 2002)

$$E(Y) = E(E(Y \mid \mu)) = E(\mu) = \exp\left(\zeta + \tfrac{1}{2}\sigma^2\right).$$

The unconditional variance is equal to the sum of the lognormal variance and Poisson variance (which for Poisson data equals the mean):

$$V(Y) = E(V(Y \mid \mu)) + V(E(Y \mid \mu)) = E(\mu) + V(\mu)$$

$$= \exp\left(\zeta + \tfrac{1}{2}\sigma^2\right) + [\exp\left(2(\zeta + \sigma^2)\right) - \exp\left(2\zeta + \sigma^2\right)]. \qquad (1)$$

Again, the unconditional variance is made up of two pieces, the Poisson variation and Poisson mean variation. The additional $\frac{1}{2}\sigma^2$ quantity in the Poisson variance (or first piece of the equation) arises from the specification of the lognormal distribution given the mean and variance of the underlying normal distribution rather than specifying the mean and variance of the lognormal distribution directly.

2.3. Underdispersed count data models

Underdispersion, where the variance of the counts is less than the mean, appears to occur less frequently and options for modeling underdispersion are not as well known as for overdispersed data. For overdispersed count data allowing the Poisson mean to be random was a standard approach to generate extra variation; this approach does not work for underdispersed count data because a random variable will add variability but does not reduce variability which is necessary for underdispersed data. Ridout and Besbeas (2004) present a nice summary and comparison of many underdispersed count models.

2.3.1. The double Poisson distribution

Originally developed to introduce a second parameter into the Poisson distribution to allow the variance to be controlled separately from the mean, the double Poisson distribution (Efron, 1986) can be used to model underdispersed count data. The probability of a given count, y, has the mass:

$$\mathrm{P}(Y=y) = \frac{\theta^{1/2} e^{-\theta\mu} e^{-y} y^{y} \left((\mu + y)/y \right)^{\theta y}}{Wy!}, \qquad y = 0, 1, 2, \ldots; \mu, \theta > 0$$

where W is a normalizing constant to ensure the distribution probabilities sum to one. If $\theta = 1$, the Poisson distribution results, while $\theta > 1$ produces underdispersed data and $\theta <$

1 generates overdispersed data. The distribution is known to have an approximate mean value of μ, an approximate variance of μ/θ, and W near one.

2.3.2. The Poisson polynomial distribution of order p

Weighted Poisson distributions are a group of distributions that are capable of accommodating both excess and lesser amounts of variability relative to the Poisson distribution. The first weighted distribution was suggested by Cameron and Johansson (1997) who suggested polynomial weights such that ϖ_y equals the weights and W_p again is the normalizing constant with p representing the pth-order polynomial:

$$\mathrm{P}(Y=y)=\frac{e^{-\mu}\mu^{y}\varpi_{y}}{Wy!},\ y=0,\ 1,\ 2,\ldots;\ \mu>0 \text{ where}$$

$$\varpi_{y}=\left(1+\sum_{i=1}^{p}\alpha_{i}y^{i}\right)^{2},\ W_{p}=\sum_{i=0}^{p}\sum_{j=0}^{p}\alpha_{i}\alpha_{j}m_{i+j}\text{, and}$$

$m_{i+j}\equiv m_{i+j}(\mu)$, the $(i+j)$th noncentral moment of the baseline Poisson distribution. The authors suggest fast simulated annealing methods to estimate the $(p+1)$ parameters of this distribution. This model was not extended beyond cross-sectional data, but the authors identify the desire to do so.

2.3.3. Castillo and Perez-Casany's weighted Poisson distributions

Castillo and Perez-Casany (1998) also proposed a weighted Poisson distribution where the form of the weights, ϖ, and normalizing constant, W are:

$$\varpi_{y}=(y+\alpha)^{r},\ \alpha>0,\text{ and } W=e^{-\mu}\sum_{y=0}^{\infty}\frac{\mu^{y}\varpi_{y}}{y!}.$$

Underdispersion arises for this weighted distribution when $r>0$, while if $r<0$ then overdispersion occurs, and the Poisson distribution results when $r=0$. They did not

extend their method to longitudinal data; therefore serial correlation and heterogeneity were not discussed.

2.3.4. Exponentially weighted Poisson distributions

A third weighted Poisson approach is advocated by Ridout and Besbeas; the weights suggested are exponential in form and center the distribution on the mean:

$$\varpi_y = \begin{cases} e^{-\beta_1(\mu - y)}, & y \le \mu \\ e^{-\beta_2(y-\mu)}, & y > \mu \end{cases}.$$

When $\beta_1 > 0$ and $\beta_2 > 0$ then underdispersion occurs, if both β_i's are negative then this distribution is overdispersed relative to the Poisson distribution; if both β_i's equal zero then the Poisson distribution results.

2.3.5. COM-Poisson distribution

In 1962, Conway and Maxwell (1962) proposed a distribution based on a queuing model with state dependent service rates; the statistical properties of this model were studied by Shmueli, et al. (2003). The form of this model is:

$$\mathrm{P}(Y = y) = \frac{\mu^y}{(y!)^\upsilon W}, \; y = 0, 1, 2, \ldots; \; \mu > 0, \; \upsilon > 0$$

where W, again, is a normalizing factor. If $\upsilon = 1$ this distribution is equivalent to the Poisson distribution, $\upsilon > 1$ represents underdispersion, and $\upsilon < 1$ is for overdispersion. A special case arises when $\upsilon = 0$ and $0 < \mu < 1$, then the geometric distribution occurs.

2.3.6. The Faddy birth process distribution

This is the distribution used in subsequent sections to model over- and underdispersion, and is presented in some detail here. To understand this distribution, recall that in a Poisson process, in the limit as $\delta t \to 0$ the probability of an event

occurring between the times t and $t+\delta t$ is $\lambda\delta t + o(\delta t)$; for simplicity, let t equal one. Faddy (1997) defined a rate parameter, λ (equation 2 below), which is used to calculate the probabilities for various rates and dispersion amounts for counts, $i = 0,1,2,\ldots,n$, where n is the maximum count possible.

$$\lambda_i = a(b+i)^c \tag{2}$$

The rate depends on the number of events observed previously, so if more events occur closer together in time the process speeds up resulting in the possibility of some very large counts and overdispersion, but if the number of events decreases over time then the process slows down and results in a consistent number of counts or underdispersion. A rate that is constant produces the Poisson distribution; a rate that increases linearly with the number of previous observations has a negative binomial distribution. For equation 2, the parameter a represents the rate of the process and must be positive ($a > 0$), b is a constant that initializes the process by ensuring that λ_0 is positive and also must be positive ($b > 0$), while c is the dispersion parameter with a maximum value of one, $c \leq 1$. If $c = 0$, the distribution is Poisson, as $c \to 1$, the negative binomial distribution results, for values of $c < 0$, the process will be underdispersed relative to the Poisson distribution.

The λ_i's are used to define an $(n+1) \times (n+1)$ matrix **U** comprised of a diagonal and superdiagonal.

$$\mathbf{U} = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & \cdots & 0 \\ 0 & -\lambda_1 & \lambda_1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ & & & -\lambda_{n-1} & \lambda_{n-1} \\ 0 & \cdots & & 0 & -\lambda_n \end{bmatrix} \tag{3}$$

If an absorbing state is desirable for the last count, n, the final element, $-\lambda_n$, can be replaced by a zero; this adjustment forces all the remaining probability onto n (i.e. $P(Y = n) = 1 - \sum_{i=0}^{n-1} P(Y = i)$). When exponentiated, **U** is an n+1 state, continuous time, Markov matrix. Finally, to find the probabilities, $p_i = P(Y = i)$, an (n+1) row vector is multiplied by the exponentiated **U** matrix:

$$\boldsymbol{p} = (p_0 \; p_1 \; p_2 \; \ldots \; p_n) = (1 \; 0 \; 0 \; \ldots \; 0) \exp(\mathbf{U}). \qquad (4)$$

Equation (4) simply assigns the probabilities, p_i, to the top row of the **U** matrix.

The above models for underdispersed count data may not be exhaustive, but are representative of the handful of ways to model this type of data. It is interesting to note that all underdispersed methods are also capable of accommodating overdispersed data, but the overdispersed methods in Section 2.2 strictly apply to overdispersed data. The underdispersed models described above were compared by Ridout and Besbeas with applications to cross-sectional data. The Ridout and Besbeas exponentially weighted Poisson model and the Faddy birth process model were extended to a longitudinal dataset, but possible serial correlation and heterogeneity were not discussed. The authors did suggest that the birth process model is likely most attractive in situations where the time variable directly influences the resulting counts.

2.4. Scale parameter greater than or less than one

It is important to note that there is a third approach to modeling dispersion in count data. Allowing the scale parameter to be greater than one in a generalized linear model (GLM) for overdispersed data or a scale parameter less than one for underdispersed data is another alternative. However, such models produce a quasi-

likelihood function versus a true likelihood function because they do not involve true probability distributions.

CHAPTER III

LONGITUDINAL COUNT DATA MODELS

When count data are collected repeatedly over time on a number of objects, this is referred to as longitudinal count data. It is reasonable to consider that the measurements on a subject are related and should not be treated as independent measurements. Longitudinal count data methods are designed to address this complication as well as others that arise when data are collected over time on the same set of objects. Many of the models discussed in this section have generalized models for time series of counts to longitudinal data. MacDonald and Zucchini (1997) is a good reference on models for time series of counts.

3.1. Serially correlated longitudinal count models

When successive observations in time are highly related with one another and less related with observations further apart in time, this persistence is known as serial correlation. This is often the primary complication addressed with longitudinal count models.

3.1.1. The Poisson-gamma longitudinal count model

The Poisson-gamma mixture model can be extended to longitudinal data by including serial correlation in the latent process. Jorgensen et al (1999) used this approach for their non-stationary state space model with a latent gamma Markov process. They incorporated covariates with a long-term effect via the latent process, while short-term covariates enter the model directly into the Poisson mean. Their model does not account for subject heterogeneity if it exists and does not discuss unequally spaced or missing data. Likewise, Henderson and Shimakura (2003) also suggest a conditionally-

distributed Poisson model compounded with an interval-specific random gamma frailty variable. The time-varying frailty variable allows for different levels across subjects and for correlation between interval counts within a subject. This model does not address underdispersion or the specific relationship between subject heterogeneity and serial correlation. A slightly different approach was introduced by Lambert (1996), who chose to use a Bayesian approach to extend a Poisson-gamma model for repeated count data collected at unequally spaced times. Both the gamma prior and Poisson counts are conditioned on an unobserved AR(1) process that accounts for the time since the last observation. This model did not discuss dispersion or subject heterogeneity.

3.1.2. The Poisson-lognormal longitudinal count model

Likewise, the Poisson-lognormal mixture model has been extended to longitudinal count data by accounting for serial correlation via an underlying process. Xu, Jones, and Grunwald (2007) (XJG model) conditioned observed counts on the exponential of a stationary, latent normal continuous time first-order autoregressive (CAR(1)) process which includes the serial correlation; the conditional counts follow a Poisson distribution. Their state space model incorporates covariates and the latent process in the mean of the Poisson process using a log link function. This model does not address over or underdispersion or subject heterogeneity. It is valuable to note that Xu (UCHSC dissertation, 2001) did discuss subject specific random effects in a Poisson-lognormal model, but this is not the model published in 2007 and here referred to as the XJG model. Czado and Kolbe (2007) similarly used a latent lognormal AR(1) process to describe the serial correlation of their conditionally Poisson time series sequence. The authors use Markov Chain Monte Carlo (MCMC) methods to estimate model parameters.

3.1.3. Generalized Estimating Equations (GEE)

GEE (Liang and Zeger, 1986) is another method for handling correlated data. For this method only the marginal mean and the covariance structure need to be specified and overall few assumptions are required for the model. Valid estimates of both regression parameters and standard errors are produced, but inefficient parameter estimates can result if a poor covariance structure is specified. In addition, since there is no identified probability distribution for this method, a likelihood function does not exist and simulation is not possible. This approach can handle over- and underdispersion through the scale parameter. Diagnostics have recently been developed by Hammill and Preisser (2006).

3.1.4. Transition models

An alternative to using a parameter-driven model where the latent parameter is used to introduce serial correlation into a Poisson model, as discussed in section 3.1.1 and 3.1.2, is to use lagged counts or an observation-driven model. Zeger and Qaqish (1988) introduced transition models for count data where the current observation conditioned on the last observed count, $y_t \mid y_{t-1}$, has a Poisson distribution. The mean of the Poisson distribution combines both covariates and the last observation. Their model does not address over or underdispersion or subject heterogeneity, though it would be straightforward to replace the Poisson distribution by a Negative Binomial distribution to model overdispersion.

Toscas and Faddy (2003) extended the Faddy distribution to longitudinal count data using an observation-driven approach to incorporate serial correlation. This model is capable of accommodating over- or underdispersion, but did not address unequally-

spaced data, and subject heterogeneity, while discussed, was unclear on how different subjects with the same mean were distinguishable.

The interpretation of regression parameters in observation-driven models is conditional on the previous observation, which may not always be the desired interpretation. Marginalized transition models (Heagerty and Zeger, 2000) have attempted to address this issue by combining a transition model that describes serial dependence with a marginal GLM that describes the mean response as a function of covariates.

3.1.5. Longitudinal Integer First-Order Autoregressive model (INAR(1))

The Poisson INAR(1) (McKenzie, 1986 and Al-Osh and Alzaid, 1987) process was originally proposed in time series models. The process uses binomial thinning (where $\circ$ is the binomial thinning operator) to include serial correlation and a count variable, ε_t, such that

$$Y_t = \phi \circ Y_{t-1} + \varepsilon_t, \ 0 \le \phi \le 1.$$

The distribution of ε_t is iid Poisson(μ), while $\phi \circ Y_{t-1}$ is a binomial random variable with the probability of success, ϕ, and number of trials, Y_{t-1}. The marginal distribution of Y_t is Poisson$\left(\mu/1-\phi\right)$. Bockenholt (1999) extended the Poisson INAR(1) model to longitudinal count data that can accommodate serial dependency and heterogeneity.

3.2. Longitudinal count data models with subject heterogeneity

A given longitudinal dataset can include serial correlation, subject heterogeneity, or both, as discussed by Jones (1990) for the normal case. This distinction is made here since none of the models in section 3.1 address subject heterogeneity separately from

serial correlation, possibly in part due to small sample sizes. Examples of longitudinal count data models that do incorporate subject heterogeneity are suggested by Zeger and Karim (1991) and Solis-Trapala and Farewell (2005). Zeger and Karim use a Gibbs-sampler approach to calculate the within-cluster correlation that arises from heterogeneity between clusters for a random effects GLM. Meanwhile, the Solis-Trapala and Farewell model uses a multivariate negative binomial model to handle overdispersed, unequally spaced longitudinal count data where random effects are used to model clusters. This model does not address serial correlation separately from subject heterogeneity.

A recent publication by Fahrmeir and Osuna (2006) provides the most inclusive count model to date. The authors use a Bayesian model to accommodate nonlinear, spatial and temporal effects, overdispersion and zero-inflated data. Random effects are used to fit various covariate effects including individual- or cluster-specific effects and spatially related effects. MCMC methods are used to accomplish the Bayesian inference, so this model is capable of handling missing data. In this article, serial correlation is not modeled separately from other count data complications, and underdispersion is not discussed.

Appendix A includes a summary of many of the models discussed in Chapter 3 and provides a side-by-side comparison of what complications are accommodated by the various longitudinal count models.

CHAPTER IV

A STATE-SPACE MODEL FOR UNDERDISPERSED, SERIALLY CORRELATED LONGITUDINAL COUNT DATA WITH SUBJECT HETEROGENEITY

4.1. The base model

4.1.1. The Xu, Jones, Grunwald longitudinal count model

We chose to extend the Xu, Jones, Grunwald (XJG) model because it is flexible enough to accommodate missing or unequally spaced, serially correlated longitudinal count data and it appeared possible to incorporate both subject heterogeneity and underdispersion. The basic idea of the XJG model is that a latent process exists which is a linear function of covariates plus an AR(1) error. Observations are assumed to have independent Poisson distributions conditional on the latent process, with means equal to the exponentiated latent process values at the observation times. The latent process is thought of as the state in a state space model, and the Poisson observations correspond to the observation equation. For unequally spaced observations the AR(1) errors are replaced by a continuous time AR(1) process, and the distance between values is accounted for in the state equation. The latent process assumes a correlation that decays at an exponential rate based on this distance. The model uses a state space representation and a modified Kalman recursion to calculate the likelihood. Parameters are estimated by numerically maximizing this likelihood. In addition this method avoids the multi-dimensional integration needed to compute the marginal likelihood since there are many observations per subject. This model will provide the basic framework for a longitudinal count data model, which we will generalize to include under and overdispersion.

The serially correlated XJG model for the special case of equally spaced observations at time *j* on subject *i* is:

$$y_{ij} \mid \varepsilon_{ij} \sim \text{ind Poisson}(\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \varepsilon_{ij})) \qquad (5)$$

$$\text{where } \varepsilon_{ij} = \phi\varepsilon_{ij-1} + \eta_{ij} \text{ and } \eta_{ij} \sim N(0, \sigma_\eta^2) .$$

The count observations y_{ij} are independent, conditional on the latent process ε_{ij}. The mean of the counts has a lognormal distribution and the random errors, η_{ij}'s, come from a normal distribution. Subjects are assumed to be independent and the random errors are independent of previous errors within subjects.

4.1.2. Accommodating missing or unequally spaced data

Missing or unequal spaced observations are handled in the XJG model by assuming an underlying continuous time AR(1) (CAR(1)) process. It is implemented by measuring the distance from the last observation and raising the correlation from the CAR(1) process to the power of this distance. Recall, the autoregressive process assumes that the correlation decays at an exponential rate based on the distance between observations. Therefore, $\phi(\delta t) = \exp(-\alpha_0(\delta t))$. For numerical optimization, a further transformation is used where $\alpha_0 = \exp(u_0)$ and u_0 is a constant parameter. This transformation forces α_0 to be positive ($\alpha_0 > 0$) which is necessary for parameter stability and results in $\phi(\delta t)$ ranging between zero and one ($0 < \phi(\delta t) \leq 1$). This autoregressive parameter is then used in the calculation of $\mathbf{Q}(\delta t)$ which is discussed in the next section.

4.1.3. The modified Kalman Filter recursion

Using the methodology employed by Xu et al., their state space representation that employs the Kalman recursion replaces the likelihood recursion step for normally distributed observations with Gaussian Quadrature used to approximate the integrals over the normal random effects, which cannot be done in closed form for Poisson observations. Calculation of the likelihood in this recursive way avoids the high dimensional integrals needed for the likelihood calculation. The likelihood is maximized to estimate the parameters when there is first-order autoregression, and can be extended, as in Section 4.2.1, if between subject variability exists.

For unequally spaced data, the length of the time step is denoted by δt. The eight step Kalman Recursion (Jones, 1993) for normal data starting with the state estimate at time t-δt, $\mathbf{s}(t\text{-}\delta t \mid t\text{-}\delta t)$, and moving to time t, $\mathbf{s}(t \mid t)$, is defined next. To simplify notation, the recursion is written for one subject and since subjects are assumed to be independent the log-likelihoods are summed over subjects and there are not any nonrandom inputs to the state or observation equations. In the basic model described above, the state is a scalar, but when subject heterogeneity is added in Section 4.2.1 it becomes a vector, so the recursion is given in vector form here.

1. Determine the prediction of the state by pre-multiplying the state transition matrix (Φ) by the current state estimate:

$$\mathbf{s}(t \mid t\text{-}\delta t) = \Phi(t;t\text{-}\delta t)\, \mathbf{s}(t\text{-}\delta t \mid t\text{-}\delta t).$$

2. Determine the prediction's covariance matrix (**P**) by pre- and post-multiplying the current covariance matrix by the transition matrix and adding the covariance matrix, $\mathbf{Q}(\delta t)$, of $\eta(\delta t)$, over the time interval difference:

$$\mathbf{P}(t \mid t\text{-}\delta t) = \mathbf{\Phi}(t;t\text{-}\delta t)\ \mathbf{P}(t\text{-}\delta t \mid t\text{-}\delta t)\ \mathbf{\Phi}'(t;t\text{-}\delta t) + \mathbf{Q}(\delta t)$$

where $\mathbf{Q}(\delta t) = \sigma_\eta^2 = \sigma_\varepsilon^2\left(1 - \phi(\delta t)^2\right)$ and (6)

$\phi(\delta t)$ is defined in section 4.1.2.

3. Determine the expected value of the next observation by pre-multiplying the predicted state by the **H** matrix. When the outcome is normally distributed, **H** is fixed.

$$\mathbf{y}(t \mid t\text{-}\delta t) = \mathbf{H}(t)\ \mathbf{s}(t \mid t\text{-}\delta t).$$

4. Calculate the difference between the observation and the expectation, which is known as the innovation in Kalman recursion:

$$\mathbf{I}(t) = \mathbf{y}(t) - \mathbf{y}(t \mid t\text{-}\delta t).$$

5. Determine the innovation's covariance matrix by pre- and post-multiplying the **H** matrix times the prediction covariance matrix and adding the observational error covariance matrix, $\mathbf{R}(t)$:

$$\mathbf{V}(t) = \mathbf{H}(t)\ \mathbf{P}(t \mid t\text{-}\delta t)\ \mathbf{H}^{\mathrm{T}}(t) + \mathbf{R}(t).$$

6. Sum the likelihood pieces across all subjects and observations.

$$\text{Sum} \leftarrow \text{Sum} + \mathbf{I}^{\mathrm{T}}(t)\mathbf{V}^{-1}(t)\ \mathbf{I}(t) \text{ and } \Delta \leftarrow \Delta + \ln|\ \mathbf{V}(t)\ |$$

where $\leftarrow$ specifies that the left side is updated by the right side for computer programming.

7. The state vector estimate is updated by taking the predicted state and adding a term which is the prediction covariance matrix pre-multiplied by the observed state elements, which is then multiplied by the inverse of the innovations covariance and the innovation:

$$\mathbf{s}(t \mid t) = \mathbf{s}(t \mid t\text{-}\delta t) + \mathbf{A}^{\mathrm{T}}(t)\ \mathbf{V}^{-1}(t)\mathbf{I}(t) \text{ where } \mathbf{A}(t) = \mathbf{H}(t)\ \mathbf{P}(t \mid t\text{-}\delta t).$$

8. Finally, the state covariance matrix is updated by taking the predicted covariance and subtracting off the innovations covariance after it is pre- and post-multiplied by **A**:

$$\mathbf{P}(t \mid t) = \mathbf{P}(t \mid t\text{-}\delta t) - \mathbf{A}^{\mathrm{T}}(t)\,\mathbf{V}^{-1}(t)\mathbf{A}(t).$$

For Poisson outcomes, the modifications to the 8 step Kalman recursion are as follows:

1. No change.

2. No change.

3. Same equation, but now $\mathbf{H}(t) = \mu(t)$, where $\mu(t)$ is the mean at time t. The **H** matrix is derived for Poisson data in Xu's dissertation (UCHSC, 2001) using first order Taylor Series expansion of the mean evaluated when each random effect equals zero.

4. No change.

5. For a count outcome the variance (equal to the mean for Poisson observations) is added at time t in place of observational error variance ($\mathbf{R}(t)$) for normally distributed data. Observational error is inherently built into the XJG model for count data since only integers can occur so the state is not observed directly but rather a Poisson observation with that mean is observed. Therefore,

$$\mathbf{V}(t) = \mathbf{H}(t)\,\mathbf{P}(t \mid t\text{-}\delta t)\,\mathbf{H}^{\mathrm{T}}(t) + \mu(t)$$

or the variance of a Poisson-lognormal

$$\mathrm{V}(Y) = \exp\left(x'\beta + \tfrac{1}{2}\sigma^2\right) + [\exp\left(2\left(x'\beta + \sigma^2\right)\right)\text{-}\exp\left(2x'\beta + \sigma^2\right)]$$

can also be used as the XJG model did.

6. Gaussian Quadrature is a numerical integration approximation method that can be used when closed form integration is not possible. This process requires that at each

observation time, the probabilities for the distribution of the outcome conditioned on random effects and the distribution of the random errors be calculated for n points across the sample space of the random errors. Multiplying these and summing as below gives the unconditional distribution of the observations, with the random errors integrated out. The likelihood pieces are then summed together over all observations and then over all subjects to get the total likelihood across all subjects and observations. The exact integral for the likelihood contribution for subject i with n_i observations is:

$$f(y_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_i \mid \varepsilon_i) g(\varepsilon_i)\, d\varepsilon_{in_i} \ldots d\varepsilon_{i2} d\varepsilon_{i1}$$

where $y_i \mid \varepsilon_i \sim \text{Poisson}\left(\mu_{ij}\right)$ and $g(\varepsilon_i)$ is a multivariate normal density.

To avoid the n_i dimensional integral, $g(\varepsilon_i)$ is expanded using its Markov property into a product of one dimensional conditional distributions:

$$g(\varepsilon_i) = g\left(\varepsilon_{in_i} \mid \varepsilon_{in_i-1}\right) \ldots\ g\left(\varepsilon_{i2} \mid \varepsilon_{i1}\right) g\left(\varepsilon_{i1}\right) \text{ where}$$

$$\varepsilon_{ij} \mid \varepsilon_{i,j-1} \sim \text{Normal}(\phi\varepsilon_{i,j-1},\ \sigma_\eta^2) \text{ and } \varepsilon_{i1} \sim \text{Normal}\ (0,\ \sigma_\varepsilon^2).$$

7. No change.

8. No change.

The likelihood that results from the Kalman recursion applied across all subjects and observations can then be optimized. The purpose of optimizing a likelihood function is to find the maximum likelihood estimates for the model parameters. A nonlinear optimizer is used for count data.

4.2. Extending the Xu, Jones, Grunwald model

The starting model for serially correlated longitudinal count data will now be extended to accommodate both subject heterogeneity and underdispersion.

4.2.1. Adding subject heterogeneity to the model

By adding between subject variability to the existing model we are allowing some subjects to have higher rates of counts while other subjects have lower rates. Figure 3 is included to provide a visual depiction of subject heterogeneity. Subjects 345's measurements (dotted line) are high relative to subject 349's measurements (solid line). The average number of inhaler usages per day for all subjects was 1.17, so typically subject 345's measurements are slightly higher relative to the population mean (subject 345's overall mean was 1.40), while measurements for subject 349 tend to run slightly below the overall mean (subject 349's overall mean was 0.52).

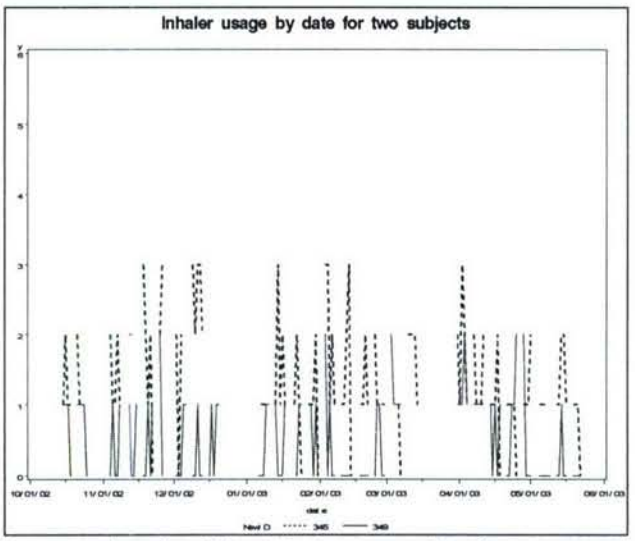


Figure 3. Graphical example of subject heterogeneity for two asthma study subjects.

The addition of subject heterogeneity to the XJG model requires adding an additional subject specific random parameter, γ_i, to the state. Since subject heterogeneity describes variation between subjects, this random effect only varies across subjects and not within a subject. Consequently, for subject *i* at time *j*, the observed counts and random effects combine for the joint distribution:

$$f(y_{ij},\gamma_i,\varepsilon_{ij}) = f(y_{ij} \mid \gamma_i,\varepsilon_{ij})f(\gamma_i,\varepsilon_{ij})$$

where the γ_i and ε_{ij}, are independent. The bivariate state vector has components, $s_{1,ij} = \varepsilon_{ij}$ and $s_{2,ij} = \gamma_i$.

In the Kalman recursions, the estimated state at a given time for a given subject also becomes bivariate, with the first level of the state equal to estimated random error, $s_1(t \mid t\text{-}\delta t) = \varepsilon_{ij}(t \mid t\text{-}\delta t)$, and the second level of the state equal to the estimated subject random effect, $s_2(t \mid t\text{-}\delta t) = \gamma_i(t \mid t\text{-}\delta t)$. The joint distribution of the estimates of the random errors, $s_1(t \mid t\text{-}\delta t)$, and random effects, $s_2(t \mid t\text{-}\delta t)$, at a given time, t_j, is assumed multivariate normal:

$$f(\varepsilon_{ij}(t \mid t\text{-}\delta t), \gamma_i(t \mid t\text{-}\delta t)) = \prod_{j=1}^{J} \frac{\exp\left(-\frac{1}{2}\boldsymbol{w}^T \boldsymbol{P}^{-1} \boldsymbol{w}\right)}{(2\pi)^{J/2}|\boldsymbol{P}|^{1/2}}.$$

Numerical integration approximation is used to find the marginal distribution of the observed data by integrating out the random effects. For the model considered in this paper, J = 2, so $\boldsymbol{w}$ is a 2 × 1 vector of the random effects centered on their mean and **P** is a 2 × 2 covariance matrix. Similar adjustments to the modified Kalman filter recursion are required when subject heterogeneity is added to the model. The state, previously a scalar, now becomes a 2 × 1 vector, while the prediction covariance matrix changes from a scalar to a 2 × 2 matrix.

The following necessary modifications to the steps of the Kalman recursion are:

1. $$\mathbf{s}(t \mid t\text{-}\delta t) = \begin{bmatrix} \varepsilon_{ij}(t \mid t-\delta t) \\ \gamma_i(t \mid t-\delta t) \end{bmatrix} = \begin{bmatrix} \phi(t; t-\delta t) & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_i(t-\delta t \mid t-\delta t) \\ \gamma_i(t-\delta t \mid t-\delta t) \end{bmatrix}.$$

2. $\mathbf{P}(t \mid t\text{-}\delta t) =$

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} \phi(t; t-\delta t) & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} \phi(t; t-\delta t) & 0 \\ 0 & 1 \end{bmatrix}^T + \begin{bmatrix} \sigma_\eta^2 & 0 \\ 0 & 0 \end{bmatrix}$$ where σ_η^2 is

defined by equation (6). The P_{kl}'s on the right hand side are all with respect to $(t-\delta t \mid t-\delta t)$ and P_{kl}'s on the left hand side are all $(t \mid t-\delta t)$, but this notation is omitted for simplicity.

3. $\mathbf{y}(t \mid t\text{-}\delta t) = [\mu(t) \quad \mu(t)z(t)]\begin{bmatrix}\varepsilon_{ij}(t \mid t-\delta t)\\ \gamma_i(t \mid t-\delta t)\end{bmatrix}$ where $\mathbf{H}(t) = [\mu(t) \quad \mu(t)z(t)]$ and $\mu(t)$ is the mean at time t and $z(t)$ represent the random subject effect design matrix at time t, which is the same as with the Poisson distribution. In our model, for subject i,

$$z(t) = \begin{cases} 1 & subject = i \\ 0 & otherwize \end{cases}.$$

4. No change.

5. $\mathbf{V}(t) = [\mu(t) \quad \mu(t)z(t)]\begin{bmatrix}P_{11} & P_{12}\\ P_{21} & P_{22}\end{bmatrix}\begin{bmatrix}\mu(t)\\ \mu(t)z(t)\end{bmatrix} + \mu(t).$

6. Gaussian Quadrature now expands from integrating across one dimension at a time to integrating across two dimensions at a time requiring a double summation to find the pieces of the likelihood within a subject that are then summed across all subjects to get the total likelihood.

7. The updated state estimate remains a 2 × 1 vector.

8. The updated state covariance estimate remains a 2 × 2 matrix.

These recursions using the bivariate Gaussian quadrature were first programmed for normally distributed outcomes and the results compared with the standard Kalman Filter recursions to ensure that the results were identical, as they should be for normal outcomes. Simulation was used to verify that subject heterogeneity was added correctly to the conditionally Poisson model of Xu, Jones, and Grunwald. The simulation required

adding subject heterogeneity to the exponentiated mean of the Poisson distribution, then randomly selecting a Poisson count given the mean. Examples of two large datasets that are simulated then optimized are shown below. Data were simulated for 100 subjects each with 100 observations and then the dataset was optimized to see how closely the parameter values used for simulation compared with the estimated parameter values; results are displayed in Table 1.

Parameter	Truth	Sim 1 Estimates	Sim 2 Estimates
$\hat{\sigma}_{\varepsilon}$	0.3	0.3378	0.2572
$\hat{\sigma}_{\gamma}$	0.2	0.2041	0.1576
$\hat{\phi}$	0.8	0.7639	0.8309
$\hat{\beta}_0$	-0.5	-0.5396	-0.4763
$\hat{\beta}_1$	0.8	0.8561	0.7689

Table 1. Results comparing the true parameters with estimated parameters based on simulated Poisson data with serial correlation and subject heterogeneity.

4.2.2. Adding underdispersion to the model

A highly flexible distribution is required to handle underdispersion (as well as overdispersion), therefore the Faddy distribution (1997) discussed in section 2.3.6 was selected to replace the Poisson distribution in the Xu, Jones, Grunwald model. The Faddy distribution was chosen for its ability to accommodate underdispersed count data. In addition, unlike most of the methods discussed in Section 2.3 for underdispersed data, a normalizing constant is not required to make the distribution a true probability distribution. Table 2 that follows demonstrates that the Faddy model produces results closer to the observed counts than the Poisson distribution does for underdispersed subjects. Recall from Section 4.2.1 subject 345 had a mean of 1.40 inhaler uses per school day, while subject 349 averaged 0.52 inhaler uses per school day.

	Subject 345			Subject 349		
y	Observed Proportion	Poisson Probabilities	Faddy Probabilities	Observed Proportion	Poisson Probabilities	Faddy Probabilities
0	0.086207	0.247449	0.086723	0.529412	0.596506	0.532438
1	0.517241	0.345575	0.510482	0.420168	0.308195	0.4201583
2	0.310345	0.241307	0.329284	0.050420	0.079617	0.045768
3	0.086207	0.112332	0.066949	0	0.013712	0.0016099
4	0	0.03922	0.00623	0	0.001771	0.0000256
5	0	0.010954	0.000321	0	0.000183	2.22E-07
6	0	0.003163	1.05E-05	0	1.7002E-05	1.17E-09

Table 2. Comparison of the observed proportions of counts, Poisson probabilities based on the mean, and Faddy probabilities based on the estimated *a* and *c* and fixed b=1.

The Faddy distribution also had previously been extended to longitudinal data by Toscas and Faddy (2003) as discussed in section 3.1.4. The approach suggested in this dissertation differs from Toscas and Faddy by using a parameter-driven approach for serial correlation versus an observation-driven approach which allows for easier interpretation of the parameters. Also, subject heterogeneity has been added and is distinguishable from serial correlation and missing or unequally-spaced data can now be accommodated.

Like the Poisson distribution, covariates will be included via the mean for the model using a log link. For subject i at time j, the equally spaced serially correlated count model with subject heterogeneity has a Faddy distribution conditioned on random effects γ (for heterogeneity) and errors ε (for serial correlation). Faddy(μ, c) represents a Faddy distribution as in Section 2.3.6 with mean μ, dispersion parameter c, and the starting parameter b set equal to 1. More detail on calculating the mean of the Faddy distribution is included in equation (9) below, and detail on constraining $b = 1$ is included in Section 4.3.1 below.

$$y_{ij} \mid \gamma_i, \varepsilon_{ij} \sim \text{ind Faddy}(\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\gamma_i + \varepsilon_{ij}), \text{c}) \qquad (7)$$

where $\gamma_i \sim N(0, \sigma_\gamma^2)$, $\varepsilon_{ij} = \phi\varepsilon_{ij-1} + \eta_{ij}$ and $\eta_{ij} \sim N(0, \sigma_\eta^2)$

This model assumes that subjects are independent, and that the γ's and ε's are independent, and the η's are independent of previous ε's and η's. The ε's within a subject follow a Gaussian first order autoregressive (AR(1)) process with variance, $\sigma_\eta^2 = \sigma_\varepsilon^2(1-\phi^2)$. Inherent in an AR(1) process is a Markov property such that only the current measurement for a subject is useful in determining where the process will be at the next time measurement.

Since a birth process does not have a simple distribution, replacing the Poisson distribution required programming the Faddy distribution. Like the conditionally Poisson mean, the Faddy mean is calculated by exponentiating the sum of the fixed effects multiplied times their parameter estimates, the random effects times their parameter estimates, and random error (equation (8)),

$$E(y_{ij} \mid \gamma_i, \varepsilon_{ij}) \approx \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\gamma_i + \varepsilon_{ij})). \quad (8)$$

Once this mean is calculated it is then used to find a, a parameter needed for the calculation of the λ_i's which are used in the Faddy distribution. From Toscas and Faddy (2003),

$$a = \frac{(\mu + b)^{(1-c)} - b^{(1-c)}}{(1-c)} \quad (9)$$

where the mean is calculated from the data using equation (8), b is fixed at one (as discussed further in section 4.3.1), and c is an estimated parameter. Once a is calculated, then the λ_i's can be computed based on equation (2), and put into the **U** matrix (defined in (3)). Once **U** is defined it can be exponentiated and the corresponding Faddy probabilities for Y given a, b, and c result from equation (4).

In order to modify the Kalman Filter recursions, some properties are needed. The marginal mean of the *y*'s can be found by using the law of total expectation along with properties of the lognormal distribution:

$$E(Y) = E[E(Y \mid \varepsilon_{ij}, \gamma_i)] \approx E\left(\exp(x'\beta + \gamma_i + \varepsilon_{ij})\right) \approx \left(\exp(x'\beta + {}^{(\sigma_\varepsilon^2 + \sigma_\gamma^2)}/_2\right). \qquad (10)$$

Likewise, the marginal variance of the y's can be approximated using the conditional variance identity along with the approximate mean of the Faddy distribution given above:

$$V(Y) = E[V(Y \mid \varepsilon_{ij}, \gamma_i)] + V[E(Y \mid \varepsilon_{ij}, \gamma_i)]$$

$$\approx E\left(\frac{\left(\exp(x'\beta + \gamma_i + \varepsilon_{ij} + b)\right)\left[1 - \left(\frac{\left(\exp(x'\beta + \gamma_i + \varepsilon_{ij})\right)}{b} + 1\right)^{2c-1}\right]}{1-2c}\right) + V\left(\exp(x'\beta + \gamma_i + \varepsilon_{ij})\right)$$

$$\approx \frac{\left(\exp(x'\beta + {}^{(\sigma_\varepsilon^2 + \sigma_\gamma^2)}/_2) + b\right)}{1-2c} * E\left[1 - \left(\frac{\exp(x'\beta + \gamma_i + \varepsilon_{ij})}{b} + 1\right)^{2c-1}\right]$$
$$+ \exp\left(2(x'\beta + \sigma_\varepsilon^2 + \sigma_\gamma^2)\right) - \exp\left(2(x'\beta) + \sigma_\varepsilon^2 + \sigma_\gamma^2\right)$$

$$\approx \frac{\left(\exp(x'\beta + {}^{(\sigma_\varepsilon^2 + \sigma_\gamma^2)}/_2) + b\right)}{1-2c} * \left[1 - \left(\frac{\exp(x'\beta)}{b} + 1\right)^{2c-1}\right] +$$
$$\exp\left(2(x'\beta + \sigma_\varepsilon^2 + \sigma_\gamma^2)\right) - \exp\left(2(x'\beta) + \sigma_\varepsilon^2 + \sigma_\gamma^2\right). \qquad (11)$$

The delta method was used to approximate

$$E\left[1 - \left(\frac{\left(\exp(x'\beta + \gamma_i + \varepsilon_{ij})\right)}{b} + 1\right)^{2c-1}\right] \approx \left[1 - \left(\frac{\exp(x'\beta)}{b} + 1\right)^{2c-1}\right].$$

Simulation was used to verify that the delta method approximation for variance derived above performs adequately. Table 3 examines the true variance and the approximate variance based on 100,000 simulated values with a fixed value $x'\beta = 0.09531$ and a normally-distributed state with mean zero and variance 0.5625, and varying *c* listed in the

first column. The variance is a better approximation for underdispersed data (when c is negative), which is well-suited for the asthma data which is known to be underdispersed. The calculated variance tends to underestimate the truth slightly and this difference grows as the data become more underdispersed.

	Truth	$[\exp(x'\beta + state)+1]^{2c}$
c	**Variance**	**Expected variance**
0	1	1
0.5	2.4561527	2.099999802
1	7.6577211	4.409999169
-0.1	0.853501	0.862097031
-0.25	0.6804792	0.690065592
-0.4	0.5495969	0.552363022
-0.5	0.4786995	0.476190521
-0.65	0.3929647	0.381166716
-0.75	0.3467688	0.328602694
-0.9	0.288499	0.263030035
-1	0.2571737	0.226757412
-1.15	0.2174643	0.181507977
-1.25	0.1946818	0.156477488
-1.5	0.1500395	0.10797973
-1.75	0.1171514	0.074513097
-2	0.0933252	0.051418924

Table 3. Simulation results to check the final piece of the model variance

The Faddy distribution results in changes to two steps of the Kalman recursion. The first change is in the calculation of the covariance matrix of the innovation or Step 5. The Poisson-lognormal variance from equation (1) was replaced by the approximate Faddy-lognormal variance derived above resulting in equation (11).

Step 6, or Gaussian Quadrature numerical integration, was also modified to handle underdispersion. This required replacing the Poisson distribution with the Faddy distribution such that

$$y_{ij} \mid \gamma_i, \varepsilon_{ij} \sim \text{Faddy}\ ((\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\gamma_i + \varepsilon_{ij}), c).$$

It was verified using the asthma inhaler data that the Faddy model with c set to 0 matched the Poisson model, and it did.

4.3. Challenges and resolutions

As with most research, difficulties are encountered and a resolution is required to move forward towards the objective. This dissertation was no different, so the major issues encountered were:

4.3.1. Fix the Faddy parameter b

Due to the time it took to optimize the Faddy distributed, serially correlated count model, we chose to fix the parameter b at one (b = 1). We still needed to add subject heterogeneity to the model and knew the optimization time might become prohibitive. Recall, for counts $i = 0,1,2,\ldots,n$ there is a rate parameter, $\lambda_i = a(b+i)^c$, in the Faddy model, where b is a constant that initializes the process by ensuring that λ_0 is positive ($b > 0$). Examination of the likelihood surface showed that it was quite flat in the region of interest when b was varied, indicating that probabilities were not sensitive to the value of b.

Y	**Sub 143** P(Y | $\hat{a}, \hat{b}, \hat{c}$)	**Sub 143** P(Y | $\hat{a}, b=1, \hat{c}$)	**Sub 345** P(Y | $\hat{a}, \hat{b}, \hat{c}$)	**Sub 345** P(Y | $\hat{a}, b=1, \hat{c}$)	**Sub 349** P(Y | $\hat{a}, \hat{b}, \hat{c}$)	**Sub 349** P(Y | $\hat{a}, b=1, \hat{c}$)
0	0.08326	0.0780572	0.0867761	0.086723	0.5325656	0.532438
1	0.3487568	0.3400037	0.510379	0.5104822	0.4201589	0.4201583
2	0.3564675	0.3562809	0.329446	0.3292844	0.0456538	0.045768
3	0.161902	0.169258	0.0668729	0.0669487	0.0015964	0.0016099
4	0.0417427	0.0466632	0.0061981	0.00623	0.0000252	0.0000256
5	0.0069741	0.0085067	0.0003176	0.0003212	2.1572E-7	2.2204E-7
6	0.00082	0.0011113	0.0000101	0.0000103	1.1217E-9	1.1706E-9
7	0.0000718	0.0001099	2.1579E-7	2.2214E-7	3.832E-12	4.062E-12
8	4.8755E-6	8.5469E-6	3.2719E-9	3.406E-9	9.106E-15	9.822E-15
9	2.6486E-7	5.3896E-7	3.674E-11	3.873E-11	1.571E-17	1.728E-17

Table 4. Comparison of probabilities for the random variable *Y* for the optimized *b* (first column for each subject) versus *b* fixed at one (second column for each subject).

Table 4 is included to show that fixing b at one has minimal impact on the probabilities; the first column for each subject represent the fitted Faddy probabilities, estimated by simulated data for the subject when all parameters were optimized over. The second column for each subject is based on simulated data for each subject when parameters a and c were optimized over and $b = 1$.

4.3.2. The assumption of log linearity

The mean of the Faddy distribution is not known exactly, however an approximation is given in Faddy (1997), which we used (equation (8) above). The assumption of log linearity between the mean and covariates was confirmed by Figure 4 below, which plots the log of the true mean, $\ln(\mu)$, versus the log of the mean calculated by the probabilities given a, b =1, and c calculated using the formula $\sum y * P(Y = y)$. This was done for several values of a, b=1, and c found by optimizing over an individual subject's results. Perfectly log-linear results would fall on the solid line; our three underdispersed subjects fell slightly above the line and the subject with the smallest mean (subject 349) deviating the furthest from the truth. Still, in practical terms we regarded these differences as small.

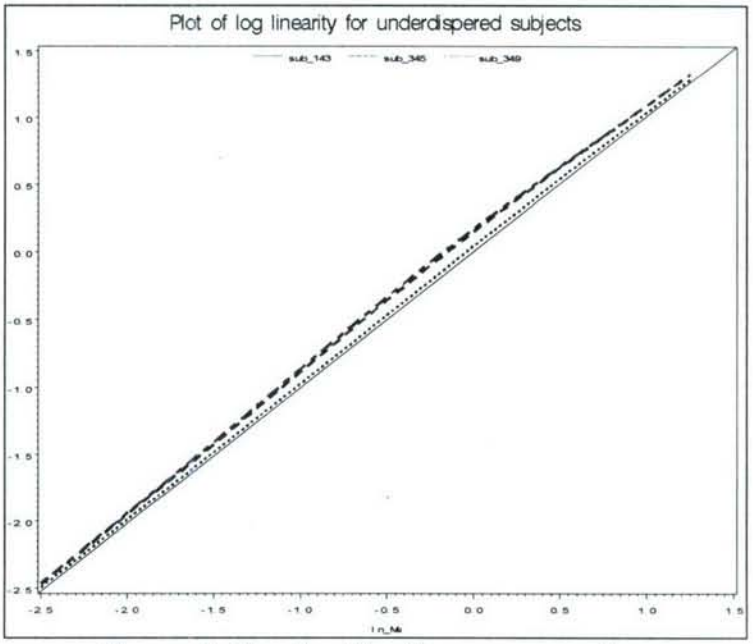


Figure 4. Check of log linear assumption

CHAPTER V

DATA ANALYSIS

5.1. Data and methods

The NJMRC asthma inhaler usage dataset has 54 subjects with between 36 and 125 observations each. The outcome of interest is the observed daily inhaler usage count which ranged from zero to six. All subjects have some missing values due to weekends, missed school days, vacations, etc. Rabinovitch, Strand, and Gelfand (2006) first studied these data to establish whether exposure to air pollution had instantaneous (i.e. same day) effects on asthma control in children with chronic illness. They used Poisson regression to model the inhaler usage data using an AR(1) covariance structure to incorporate serial correlation in a GEE model. Underdispersion was taken into account with the scale parameter. Missing data were accounted for correctly by the AR(1) structure, but subject heterogeneity was not incorporated into the analysis.

The model developed in chapter 4 is now used to analyze the data. We primarily focused on the underdispersed group of subjects since the majority of the subjects are represented by that large cluster (see Figure 2); so data for 48 subjects are included in the results that follow. Time-varying covariates that were used in this analysis, which included some of the covariates considered by Rabinovitch, et al, are: barometric pressure (press), temperature (temp), and humidity (hum), an air pollution variable (morning maximum particulate matter 2.5 which is denoted mmaxpm25), and a physical activity indicator variable (PA). The indicator for physical activity equaled one if gym class was not scheduled for that day and zero for days with gym class. In a typically week, gym class was held Monday through Thursday, so the indicator can also be thought

of as a Friday indictor with one representing Fridays which were days without gym class. In addition, all models include the following parameters: fixed effects regression parameters β_0 for level and β_1 for linear time trend, c for dispersion in the Faddy distribution, ϕ for the correlation coefficient of the latent AR(1) process, σ_ε^2 for the variance of the continuous time latent AR(1) process, and σ_γ^2 for the variance of subject heterogeneity. Covariates were scaled to eliminate the exponentiation of very large values, as follows: morning maximum particulate matter 2.5 (mmaxpm25) = mmaxpm25/10, barometric pressure (press) = press/1000, temperature in °F (temp) = temp/100, and humidity (hum) = hum/100.

5.2. Results

5.2.1 Model selection

Table 5 contains results from a forward-step regression methodology to study the association of the added independent variables with inhaler use.

Model description	# of estimated parameters	-2LL	AIC
Model 1: intercept & time (i.e. time trends)	6	12612.7376	12624.7376
Model 2: intercept, time & physical activity indicator (i.e. PA)	7	10706.33897	10720.33897
Model 3: intercept, time, PA, & morning maximum particulate matter 2.5 (mmaxpm25)	8	10705.45546	10721.45546
Model 4: intercept, time, PA, mmaxpm25, & barometric pressure (press)	9	10704.64674	10722.64674
Model 5: intercept, time, PA, mmaxpm25, press, & temperature (temp)	10	10704.45636	10724.45636
Model 6: intercept, time, PA, mmaxpm25, press, temp, & humidity (hum)	11	***10692.47511***	***10714.47511***
Model 6A: A conditionally **Poisson** model with intercept, time, PA, mmaxpm25, press, temp, hum	10	12568.60775	12588.60775

Table 5. Longitudinal regression results for underdispersed subjects

Four additional models were considered after the forward-step regression was complete. These models are included in Table 6 below along with a description of the model, the number of estimated parameters, the -2LL, and AIC.

Model description	# of estimated parameters	-2LL	AIC
Model 7: intercept, time, PA, press, temp, & hum (no mmaxpm25)	10	10695.55074	10715.55074
Model 8: intercept, time, PA, mmaxpm25, & hum (no press or temp)	9	10697.98283	10715.98283
Model 9: intercept, time, PA, mmaxpm25, press, temp, & hum ($\phi = 0$)	10	10799.53678	10819.53678
Model 10: intercept, time, PA, & hum (no mmaxpm25, press, or temp)	8	10699.77825	10715.77825
Model 11: intercept, time, PA, mmaxpm25, press, temp, & hum ($\sigma_\gamma = 0$)	10	10747.67731	10767.67731

Table 6. Additional longitudinal regression results for underdispersed subjects

The indicator variable for physical activity is clearly the most influential covariate based on the reduction in -2LL relative to all the other covariates. Humidity appears to be the most important weather variable based on the significant increase in AIC when it was not included in the model. Both pressure and temperature on the other hand had very little impact on AIC when they were not included and mmaxpm25 also had little impact on the AIC when it was not included.

5.2.2. Model interpretation

The smallest -2 log likelihood (-2LL) is produced with the indicator variable for physical activity, the air pollution variable and all the weather variables.

$$E(Y_{ij} \mid \gamma_i = 0, \varepsilon_{ij} = 0) =$$

$$\exp\left(\beta_0 + \beta_1 t_{ij} + \beta_2 PA_{ij} + \beta_3 MMAXPM25_{ij} + \beta_4 PRESS_{ij} + \beta_5 TEMP_{ij} + \beta_6 HUM_{ij}\right) \qquad (10)$$

Parameter estimates for the model in equation 10 are included in Table 7.

Parameter Estimates	$\hat{\beta}_0$	$\hat{\beta}_{1,TIME}$	$\hat{\beta}_{2,PA}$	$\hat{\beta}_{3,MMAXPM\ 25}$	$\hat{\beta}_{4,PRESS}$	$\hat{\beta}_{5,TEMP}$	$\hat{\beta}_{6,HUM}$
	-1.18	-0.22	-1.29	0.016	2.1	-0.25	-0.23
	$\hat{\sigma}_\varepsilon$	$\hat{\sigma}_\gamma$	$\hat{\phi}$	$\hat{c}$			
	0.25	0.28	0.78	-2.25			

Table 7. Parameter estimates for Model 6.

The estimate for physical activity ($\hat{\beta}_2$ = -1.29) decreases the expected number of puffs by more than one unit which is expected since the indicator variable equals one on the days when there was not physical activity and subjects did not need to pretreat their asthma. As mmaxpm25 increases asthma inhaler uses increases ($\hat{\beta}_3$ = 0.016), however the change in AIC due to mmaxpm25 is small and indicates non-significance. Inhaler use increases as barometric pressure increases also ($\hat{\beta}_4$ = 2.1), while use decreases as temperature and humidity increase, $\hat{\beta}_5$ = -0.25 and $\hat{\beta}_6$ = -0.23 respectively. From the AIC values it appears that humidity is more important than temperature or pressure for inhaler use counts. There appears to be moderately strong serial correlation, $\hat{\phi}$ = 0.78 with relatively small variability, $\hat{\sigma}_\varepsilon$ = 0.25, while omitting phi from the model substantially increased AIC (model 9 AIC = 10819.53678). In addition, there is some variability across subjects, $\hat{\sigma}_\gamma$ = 0.28 and strong underdispersion due to the negative *c* value, $\hat{c}$ = -2.25. Setting c=0 gives a model with Poisson random variation and the AIC is much larger for the Poisson model, indicating a much better fit for the Faddy model. The most parsimonious model is model 10 which only uses four independent variables, the intercept, time trend, physical activity, and humidity. Model 10 does not produce the smallest -2LL, but the AICs for model 6 (AIC = 10714.475) and model 10 (AIC = 10715.778) are very close in value while saving three degrees of freedom.

5.2.3. Comparison with GEE and Poisson approaches

The estimates produced above were compared with the GEE approach using the GENMOD procedure in SAS® and with the conditionally Poisson model, model 6A above (the model of Xu et al. with subject heterogeneity added); standard errors for GEE are listed below the parameter estimates in parentheses in Table 8.

Model	$\hat{\beta}_0$	$\hat{\beta}_{1,TIME}$	$\hat{\beta}_{2,PA}$	$\hat{\beta}_{3,MMPM\,25}$	$\hat{\beta}_{4,PRESS}$	$\hat{\beta}_{5,TEMP}$	$\hat{\beta}_{6,HUM}$
GEE w/AR(1) covar structure	-0.9443 (1.1793)	0.0255 (0.0202)	-1.2458 (0.0841)	0.0175 (0.0069)	0.5848 (1.8636)	-0.4462 (0.1219)	-0.3093 (0.0619)
GEE w/Comp. Sym. covar structure	-0.6717 (1.1023)	0.0237 (0.0225)	-1.2208 (0.0813)	0.0153 (0.0069)	0.0716 (1.7477)	-0.3678 (0.1273)	-0.2194 (0.0639)
Model 6A	0.4828	0.02318	-1.2214	0.01494	0.1093	-0.3687	-0.2196
Model 6 (Table 7 for comparison)	-1.1822	-0.2223	-1.2939	0.01630	2.1004	-0.2527	-0.2349

Table 8. GEE and Poisson Model 6 parameter estimates.

The parameter estimates in Table 8, excluding the barometric pressure estimate, $\hat{\beta}_4$, are comparable, especially between the GEE model with compound symmetric covariance structure and the conditionally Poisson model. Barometric pressure appears quite variable resulting in the large standard error estimates. The intercept estimates, $\hat{\beta}_0$, likely differ since in the Poisson model they have a subject-specific interpretation while in the GEE approach they have a population average interpretation. The other regression coefficients can be interpreted in either manner, subject-specific or population average (Diggle et al, 2002, p. 137). Some of the Faddy estimates are slightly off from the GEE and Poisson model estimates, most notably for the time trend. The approximation to the mean results in an upwards trend for underdispersed data which causes the time trend to overestimate the true effect. To correct this problem we used a second order polynomial to better represent a as a function of the mean and c in place of equation 9.

$$a = 0.1705 + 0.0448 * c + 0.6662 * \mu + 0.1012 * c * \mu + 0.0819 * \mu^2 - 0.5237 * c * \mu^2 .$$

The resulting parameter estimates for Model 6, based on the above change to a are shown in Table 9.

Parameter Estimates after mean correction	$\hat{\beta}_0$	$\hat{\beta}_{1,TIME}$	$\hat{\beta}_{2,PA}$	$\hat{\beta}_{3,MMAXPM25}$	$\hat{\beta}_{4,PRESS}$	$\hat{\beta}_{5,TEMP}$	$\hat{\beta}_{6,HUM}$
	0.9809	0.1264	-1.2652	0.0108	-0.2454	-0.6062	-0.3478
	$\hat{\sigma}_\varepsilon$	$\hat{\sigma}_\gamma$	$\hat{\phi}$	$\hat{c}$			
	0.2275	0.4277	0.8317	-2.019			

Table 9. Parameter estimates for Model 6 after mean correction.

While the time trend parameter still appears too large relative to the GEE and Poisson estimates, it has moved towards a more reasonable value indicating that the need to better approximate the mean is the root of the problem. We are continuing to work on developing a better approximation for a.

CHAPTER VI

DIAGNOSTICS

6.1 A general method of model assessment for longitudinal data

Diagnostics provide a way to distinguish between models, and to determine if a given model is appropriate for a given dataset. In normal regression models, residuals are one of the main tools for model diagnostics. In more complex situations such as longitudinal data or non-normal outcomes, residuals become more difficult to use and interpret. Here we consider a general method for model diagnostics based on the ideas of parametric bootstrap, as discussed for time series by Tsay (1992) and Grunwald et al. (2000).

The basic process is:

(1) Fit each model under consideration to the dataset being modeled

(2) Use each of the fitted models to simulate one or more sets of counts

(3) Compare the simulated counts with the observed counts to determine the compatibility of the fitted models with the observed data

(4) Calculate 95% confidence intervals for key parameters and quantities based on several simulations of the fitted models and these results are compared with the observed values in the dataset to determine which models best capture those results.

Ideally this method would be applied to several existing longitudinal count models to rule out which models would not be appropriate for this underdispersed dataset.

It is relevant to note that this method is of limited value when studying GEE methods, since those methods do not fully specify the probability distributions and so do

not allow for simulation. Thus, the model fit of GEE cannot be assessed using the method given above. It is still possible to study the performance of GEE methods (e.g. bias of parameter estimates) by repeatedly simulating and fitting data from a model such as (10), and summarizing the results as usual to assess bias and efficiency. Therefore, even with GEE methods the models of Chapter 4 are useful in providing data for simulation studies.

6.2 Application to NJMRC asthma data

6.2.1 Considering potential models

When selecting a model for longitudinal count data analysis, it is practical to determine which models are appropriate for the dataset. By starting with a list of models, then identifying the capabilities of the model, such as in Attachment 1, you can rule out models that do not meet the requirement of the data. For example, if you having missing or unequally spaced data in your model, it would not be practical to select a model from Attachment 1 that is not capable of accommodating missing data unless you want to modify the model for this complication, which may not be feasible. This is an important step since fitting and simulating from some models may require special software that may not be easily available, so eliminating models that do not appear good candidates before fitting saves time.

Visual inspection is another tool to help determine the appropriateness of a given model. Simulation can be used to determine if data simulated from a model appear reasonable for the given dataset being modeled. Ideally each model will be fitted to get parameter estimates to be used in simulation. However, in most cases simulation is much easier than fitting and so simulating from a variety of parameter values may give

sufficient information about the model behavior. Figure 5 shows count data simulated for two subjects from a simple version of the Jorgensen (1999) model that includes subject heterogeneity and serial correlation as he does, but no covariates, and uses one set of arbitrarily chosen parameters. From these simulations, the simulated data from this model does not look anything like the asthma data plots in Figures 1 and 3.

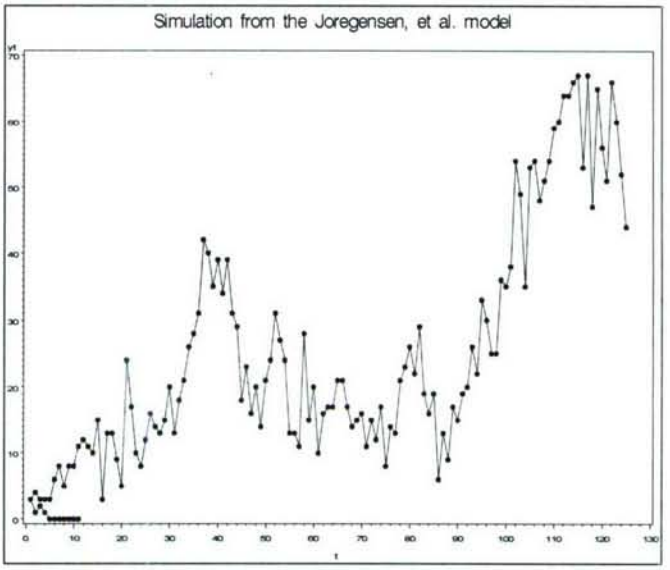
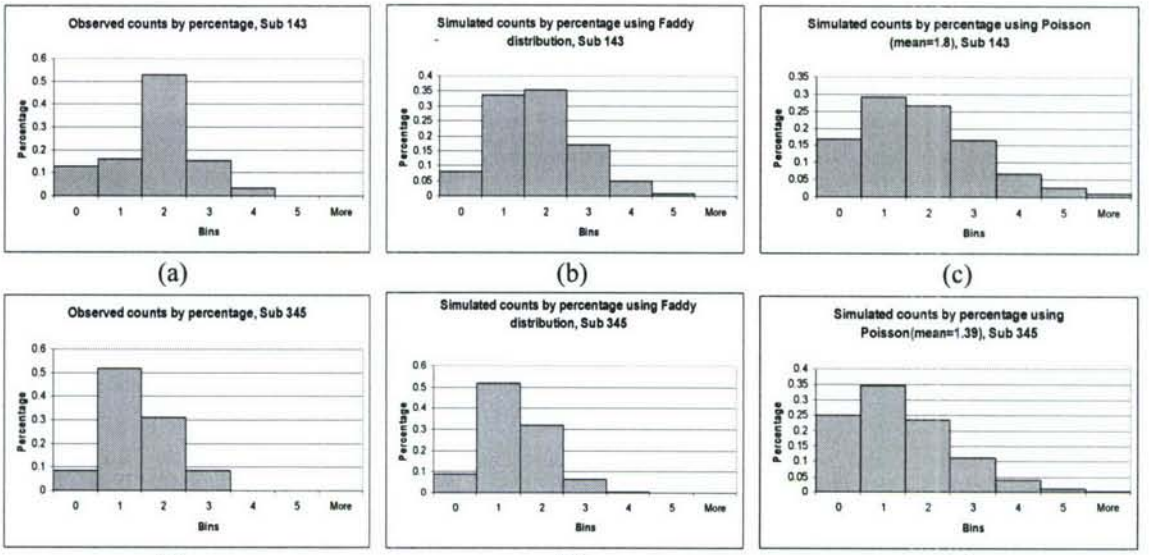


Figure 5. Data simulated for two subjects from the Jorgensen, et al model.

Visual inspection can also be used to examine other properties. In Figure 6 below, two individual subjects' observed data by percentage (a) are shown in the histograms and next to the original data, large samples were simulated for two models based on the estimated parameters *a* and *c* for the Faddy distribution, (b) and using the subject's mean as the Poisson parameter estimate (c). The models using the Faddy distribution appear to better replicate the observed data than does the Poisson distribution.



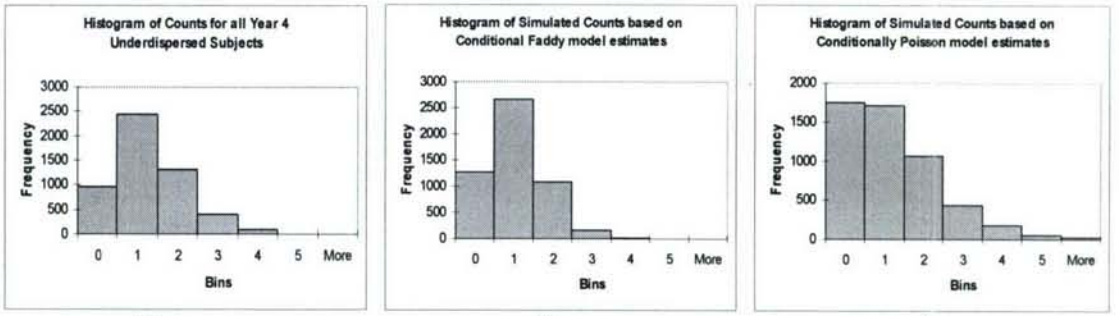
(a) (b) (c)

(a) (b) (c)

Figure 6. Comparison for two subjects of the observed data (a), with results based on simulation of the Faddy distribution (b) and the Poisson distribution (c).

6.2.2 Model diagnostics for dispersion

Using the fitted model results for model 6 in Chapter 5, inhaler use counts were simulated for 48 subjects. The simulation used the year 4 data for the physical activity indicator variable, the weather variables and the morning maximum particulate matter 2.5 variable. While not a perfect fit, the histogram in Figure 7(b) below, based on the model developed in Chapter 4 and parameter estimates given in Table 7, better matches the observed data in 7(a) than the simulated data in histogram 7(c) using the conditionally Poisson model and parameter estimates given in Table 8. This is a visual validation of the substantial decrease in likelihood noted in Table 5 for the Faddy model compared with the Poisson model.



(a) (b) (c)

Figure 7. Comparison for all 48 Year 4 underdispersed subjects observed data (a), with results based on simulation of the Faddy distribution (b) and the Poisson distribution (c).

6.2.3 Model diagnostics for serial correlation

This method of model assessment was applied to the model estimated in Chapter 5 for the NJMRC asthma data. The parameter ϕ was studied in two ways to better understand why a higher value of ϕ was consistently produced by the many models considered for the asthma data. This examination also suggested some quantities that could be calculated for the actual data and compared with values calculated from simulated data to study whether the serial correlation properties of the model captured the patterns seen in the data. ϕ describes the amount of correlation between the current observation and the previous observation, so the first method to examine ϕ was to measure the difference between the current observation and the previous, or lagged, observation (i.e. difference = y – lag(y)). For strongly correlated variables, you would expect to see smaller differences than you would see for uncorrelated variables, in analogy with the usual normal AR(1) model. Table 10 shows the results of the differences in the asthma dataset for underdispersed subjects regardless of the time distance between the two consecutive observations. The differences between the current inhaler use and previous inhaler use in the asthma data are typically zero or one; over

80% of the asthma data differ by these amounts indicating a strong relationship between successive values.

Difference	Frequency	%
-5	2	0.038752
-4	16	0.310017
-3	82	1.588839
-2	369	7.149777
-1	1073	20.790544
0	2124	41.154815
1	1025	19.860492
2	379	7.343538
3	77	1.491959
4	11	0.213137
5	1	0.019376
6	2	0.038752

Table 10. Difference in count value between y and lag(y) for underdispersed subjects.

Simulation was used to study how the differences between consecutive values ranged for different values of ϕ. The simulated model from Section 6.1.2 is again used here, where the observed values for physical activity, mmaxpm25, barometric pressure, temperature, and humidity were used to generated asthma inhaler use counts based on the conditional Faddy model in equation (10). Table 11 demonstrates that to a small degree as ϕ increases, large differences tend to occur less frequently and values tend to change in smaller increments as more probability accumulates at differences of zero, negative and positive one. The differences due to different values of ϕ are not as large as expected, but may be impacted by missing data, and the model with phi set to 0 (Model 9) in Chapter 5 had a substantially larger likelihood, indicating a model including serial correlation was an improvement. The values of ϕ were chosen to give a wide range of possible values and $\phi = 0.779$ was used since that was the model estimate in Table 7. For all of these values of phi, the percents are similar to those observed in the real data, indicating a reasonable model fit.

ϕ=0.25			ϕ=0.5			ϕ=0.779			ϕ=0.99		
Diff	**Freq**	**%**	**Diff**	**Freq**	**%**	**Diff**	**Freq**	**%**	**Diff**	**Freq**	**%**
-4	1	0.019	-4	2	0.039	-4	1	0.019	-4	2	0.039
-3	42	0.814	-3	35	0.678	-3	32	0.620	-3	36	0.698
-2	285	5.522	-2	289	5.600	-2	281	5.445	-2	280	5.425
-1	1297	25.131	-1	1280	24.801	-1	1276	24.724	-1	1260	24.414
0	1936	37.512	0	1978	38.326	0	2013	39.004	0	2019	39.120
1	1256	24.336	1	1230	23.833	1	1224	23.716	1	1239	24.007
2	302	5.852	2	310	6.007	2	300	5.813	2	290	5.619
3	40	0.775	3	36	0.698	3	32	0.620	3	33	0.639
4	2	0.039	4	1	0.0194	4	2	0.0388	4	2	0.0388

Table 11. Difference in count value between y and lag(y) for simulated data based on asthma data parameter estimates for varying values of ϕ.

Since the differences observed in Table 11 appear negligible, the simulation was redone with one change, $\hat{\sigma}_\varepsilon$ was increased from 0.25 to 0.6 (all other parameters are based on Table 7) to allow for more variability in the variance associated with ϕ. The simulations using the new standard deviation for ε, allow for a better understanding of the serial correlation variable, ϕ. In table 12, there is a decrease in the number of larger differences as ϕ increases and the data become more concentrated at differences of zero, negative and positive one, as expected.

ϕ=0.25, $\hat{\sigma}_\varepsilon$=0.6			ϕ=0.5, $\hat{\sigma}_\varepsilon$=0.6			ϕ=0.75, $\hat{\sigma}_\varepsilon$=0.6			ϕ=0.99, $\hat{\sigma}_\varepsilon$=0.6		
Diff	**Freq**	**%**	**Diff**	**Freq**	**%**	**Diff**	**Freq**	**%**	**Diff**	**Freq**	**%**
-7	4	0.08	-7	0	0	-7	0	0	-7	0	0
-6	8	0.16	-6	5	0.10	-6	1	0.02	-6	1	0.02
-5	8	0.16	-5	10	0.19	-5	8	0.16	-5	2	0.04
-4	40	0.78	-4	27	0.52	-4	29	0.56	-4	12	0.23
-3	109	2.11	-3	120	2.33	-3	81	1.57	-3	53	1.03
-2	444	8.60	-2	385	7.46	-2	371	7.19	-2	284	5.50
-1	1151	22.30	-1	1193	23.12	-1	1172	22.71	-1	1228	23.79
0	1631	31.60	0	1705	33.04	0	1853	35.90	0	2031	39.35
1	1166	22.59	1	1153	22.34	1	1139	22.07	1	1180	22.86
2	420	8.14	2	402	7.79	2	388	7.52	2	294	5.70
3	123	2.38	3	116	2.25	3	87	1.69	3	62	1.20
4	39	0.76	4	33	0.64	4	26	0.50	4	12	0.23
5	10	0.19	5	6	0.12	5	5	0.10	5	2	0.04
6	6	0.12	6	6	0.12	6	1	0.02	6	0	0
7	2	0.04	7	0	0	7	0	0	7	0	0

Table 12. Difference in count value between y and lag(y) for simulated data based on asthma data parameter estimates, excluding $\hat{\sigma}_\varepsilon$ = 0.6, for varying values of ϕ.

The second diagnostic developed to examine ϕ was to study the length of runs, or repeated count values. For higher values of ϕ you would expect to see longer runs than for smaller values of ϕ. The length of a run for underdispersed subjects in the asthma data ranged from one, meaning the current value observed was different than the previous value, to fourteen, which indicated that the same value occurred fourteen times in a row for one subject. The results for the underdispersed National Jewish subjects are summarized in Table 13.

Length of Run	1	2	3	4	5	6	7	8	9	10	11	12	13	14	...	33
# occurrences	2117	459	224	154	61	21	16	14	6	1	4	3	2	2	0	1

Table 13. The length of the run for a repeated inhaler count for one subject then results for all underdispersed subjects were combined.

Again for varying values of ϕ, simulation was used to study how the lengths of the runs change. Results for the simulation of asthma inhaler use subjects for varying ϕ values are displayed in Table 14.

length of string	ϕ=0.25	ϕ=0.5	ϕ=0.779	ϕ=0.99
1	2168	2126	2067	2110
2	649	637	654	603
3	247	252	248	231
4	116	119	134	143
5	51	48	44	48
6	21	20	25	22
7	14	19	12	15
8	5	8	8	14
9	1	0	4	1
10	1	2		3

Table 14. The length of the run for varying values of ϕ for simulated data.

Typically longer runs occur more frequently for higher values of ϕ, but as before this difference across values of ϕ is not striking. Therefore, in Table 15 below we examined the length of runs produced for simulated data based on parameter estimates

from table 7, with the single change of increasing the variability of $\hat{\sigma}_\varepsilon = 0.6$. The length of runs increases as ϕ increases very clearly and the influence of ϕ is better understood.

length of string	ϕ =0.25	ϕ =0.5	ϕ =0.75	ϕ =0.99
1	2529	2449	2228	2079
2	686	661	700	618
3	224	233	236	252
4	86	107	131	123
5	35	30	38	53
6	11	14	11	30
7	5	5	9	8
8	2	3	0	7
9		2	1	2
10			1	4
11			1	0
12				0
13				0
14				1
15				0
16				0
17				1

Table 15. The length of the run for varying values of ϕ for simulated data with $\hat{\sigma}_\varepsilon = 0.6$.

CHAPTER VII

STRENGTHS, LIMITATIONS, AND FUTURE RESEARCH

7.1. Strengths

The model discussed in Chapter 4 is designed to accommodate the most complicated longitudinal count datasets. The model is capable of handling serially correlated, over- or underdispersed or Poisson count data, while allowing for unequally spaced or missing data with fixed and random covariates, including the case of subject heterogeneity (random intercepts). In addition, the state space model produces a likelihood which can be a basis for comparison to other likelihood approaches for longitudinal count data, and the probability model can be used to simulate data with all of the above listed complexities. This would be useful for example in simulation studies of the performance of GEE methods for underdispersed longitudinal data.

7.2. Limitations

As with most research projects there are trade-offs to the strengths gained and my research is no different. The time it takes to optimize the model discussed in Chapter 4 is a major limitation with this method. Models 1 through 6 in chapter 5 took between 8 and 26 hours to run. More complicated models with additional parameters or interaction terms were not considered due to the time to optimize. We were also unable to conduct the usual simulation studies to assess bias and variability of the model parameter estimates. It is certainly possible to program the model in another software program such as FORTRAN or C, but with no experience with these software and uncertain availability of the software, reprogramming was not attempted. We have considered other means of speeding up the optimization, including replacing the Gaussian quadrature with Laplace

approximations, or attempting to calculate the Faddy probabilities outside of the loops in the likelihood. These would be worth exploring.

The nonlinear optimization routine by quasi-Newton method (NLPQN) used in SAS/IML® does not compute second-order derivatives and therefore no standard errors for the estimates can be determined, making it difficult to know how much variability exists for a parameter. With faster code, this could be addressed by using simulation to calculate the standard errors.

Optimization results did vary across subjects and while typically the Faddy distribution did better than the Poisson distribution (see Table 2), it was not always superior. Specifically, as the underdispersed subject's mean grew, the Faddy fit struggled to match the appropriate proportions of observed counts and the Poisson distribution seemed to do as well as the Faddy distribution, see Table 16.

	Subject 143		
y	Observed Proportion	Poisson Probabilities	Faddy Probabilities
0	0.128	0.165299	0.078057
1	0.16	0.297538	0.340004
2	0.528	0.267784	0.356281
3	0.152	0.160671	0.169258
4	0.032	0.072302	0.046663
5	0	0.026029	0.008507
6	0	0.010378	0.001111

Table 16. Comparison of the observed proportions of counts, Poisson probabilities based on the mean, and Faddy probabilities based on the estimated *a* and *c* and fixed b=1.

7.3. Future research

Many interesting ideas presented themselves throughout the study of longitudinal count data methods, the Faddy distribution, and the asthma inhaler use dataset. Unfortunately, every idea was not explored at the desired depth in an effort to stay

focused on the original objectives identified which remained valid throughout my research.

One of the most interesting ideas suggested was the notion of multiple dispersion parameters, c_k's, where $1 < k \leq n$, the number of study subjects. This idea makes potential sense for the asthma dataset which appears to have some clustering between subjects who pretreat their asthma prior to physical activity versus those who do not. Also, each subject could possibly have their own dispersion parameter, c_k. Multiple dispersion parameters would require different mean and variance calculations for each c_k.

Another idea that could use further development is to program many models in Attachment 1 and compare them on the same dataset. This can be quite complicated since some models can accommodate missingness or unequally spaced data, underdispersion, or subject heterogeneity, while other methods don't explain how it might be accomplished or are not applicable when some complications are present in the data. For the asthma data, none of those models can accommodate underdispersion and so none would be appropriate for the NJMRC asthma data, but in other situations more candidate models may be available.

There is also the possibility to use Markov Chain Monte Carlo (MCMC) methods to estimate parameters, which has the potential to rapidly compute the model and provide standard errors for estimates. We considered this method, but due to limited programming capabilities in WinBUGS, which currently lacks the ability to exponentiate a matrix, the appropriate code for the Metropolis-Hastings algorithm would need to be written in a programming language such as C.

APPENDIX A

COMPARISON CHART OF LONGITUDINAL COUNT DATA MODELS

	Subject Heterogeneity Incorporated?	*Parameter- or Observation-Driven Model?*	*Cond. dist: Overdispersed, Underdispersed, neither, or both?*
Longitudinal INAR(1) (Bockenholt)	Yes, via finite mixture models	Observation-driven	Conditionally underdispersed.
Jorgensen, et al Statespace model	No	Parameter-driven	Conditionally neither.
Lambert Model (Bayesian methodology)	No, in part due to small sample size.	Parameter-driven	Conditionally neither.
Poisson w/random gamma mean (Henderson & Shimakura)	Frailty variable, Z_j	Parameter-driven	Conditionally neither.
Xu, Jones, & Grunwald Model	No	Parameter-driven	Conditionally neither.
Transition Models (Zeger and Qaqish)	No, but could be modified.	Observation-driven	Conditionally neither.
Toscas & Faddy	No, but could be modified.	Observation-driven	Conditionally, either over or underdispersed.

	Subject Heterogeneity Incorporated?	*Parameter- or Observation-Driven Model?*	*Cond. dist: Overdispersed, Underdispersed, neither, or both?*
Longitudinal INAR(1) (Bockenholt)	Yes, via finite mixture models	Observation-driven	Conditionally underdispersed.
Jorgensen, et al Statespace model	No	Parameter-driven	Conditionally neither.
Lambert Model (Bayesian methodology)	No, in part due to small sample size.	Parameter-driven	Conditionally neither.
Poisson w/random gamma mean (Henderson & Shimakura)	Frailty variable, Z_j	Parameter-driven	Conditionally neither.
Xu, Jones, & Grunwald Model	No	Parameter-driven	Conditionally neither.
Transition Models (Zeger and Qaqish)	No, but could be modified.	Observation-driven	Conditionally neither.
Toscas & Faddy	No, but could be modified.	Observation-driven	Conditionally, either over or underdispersed.

	How are covariates incorporated?	*Does model address stationarity/non-stationarity?*	*Accommodate unequally spaced data/missing data?*
Longitudinal INAR(1) (Bockenholt)	Via the mean of the error variable, *et.*	Yes, stationary Poisson	No/No
Jorgensen, et al Statespace model	Short-term covariates via conditional Poisson, long-term covariates via the latent process	Yes, non-stationary	No/No
Lambert Model (Bayesian methodology)	The pdf portion using a log link function.	No	Yes/Yes
Poisson w/random gamma mean (Henderson & Shimakura)	Via the mean of the Poisson	No	Yes/Yes
Xu, Jones, & Grunwald Model	Via conditional mean of the Poisson	Yes, latent process is stationary	Yes/Yes
Transition Models (Zeger and Qaqish)	Via the mean of the Poisson using a log link.	No	No/No
Toscas & Faddy	Via the log mean of the distribution; complex relation to the rate parameter	No	No/No

	What is the interpretation of the model parameter(s)?	*Weakness(es) of model:*
Longitudinal INAR(1) (Bockenholt)	Condition on previous observations.	Difficult to estimate--efforts have emphasized stocastic processes versus estimation.
Jorgensen, et al Statespace model	Conditional on covariates.	No subject heterogeneity. Does not discuss unequally spaced or missing data.
Lambert Model (Bayesian methodology)	The mean rate for a given dosage.	No subject heterogeneity or underdispersion.
Poisson w/random gamma mean (Henderson & Shimakura)	Conditional on covariates.	Does not explain the constraint between serial correlation and subject heterogeneity.
Xu, Jones, & Grunwald Model	Conditional on covariates.	No subject heterogeniety and no dispersion addressed.
Transition Models (Zeger and Qaqish)	Predicts the count in the next interval given the count in the previous interval and covariates.	Conditional interpretation of parameters based on previous interval. Ad hoc assumptions needed to handle 0's.
Toscas & Faddy	Conditional mean based on last observation	No subject heterogeneity. Conditional interpretation. Complex method of relating covariates to parameters.

APPENDIX B

SIMULATE FADDY AR(1) COUNTS WITH SUBJECT HETEROGENEITY

```
data sim;
seed1=25585393;
phi=0.987;
s_e=0.415;
sgam=0.71;
b=1;
c=-1.94;
n_sub=10;
n_obs=100;
n=11;
fac=sqrt(1-phi**2);
s_eta=s_e*fac;
do i=1 to n_sub;
  id=i;
  if i<=n_sub/2 then x=1;
  else x=0;
  call rannor(seed1,sgam0);
  gam=sgam*sgam0;
  do  j=1 to n_obs;
  if j=1 then do;
       call rannor(seed1,s_eta0);
       u=s_eta0*s_eta;
       eps=u/fac;
    end;
    else if j > 1 then do;
       call rannor(seed1,s_eta0);
       u=s_eta0*s_eta;        ***use square root of variance---seta;
       eps=eps*phi+u;
    end;
    mu=exp(-0.8+0.5*x+eps+gam);
    a=((mu+b)**(1-c)-b**(1-c))/(1-c);
    output;
  end;
end;
run;

proc iml;
use sim;
read all var {a b c n id j seed1};
k=0;
z=nrow(a);
y=j(z,1,99);
do ii=1 to z;
  k=k+1;
  lambda=shape(1,11,1);
  lambdaall=shape(1,12,1);
  do t=1 to n by 1;
    lambda[t]=a[k,1]*(b[k,1]+t)**c[k,1];
  end;
  l_not=a[k,1]*(b[k,1])**c[k,1];
  lambdaall=l_not//lambda;
```

```
  convert={1 1 1 1 1 1 1 1 1 1 1 1};
  Qp1=lambdaall*convert;
  Qp2={-1 1 0 0 0 0 0 0 0 0 0 0,
       0 -1 1 0 0 0 0 0 0 0 0 0,
       0 0 -1 1 0 0 0 0 0 0 0 0,
       0 0 0 -1 1 0 0 0 0 0 0 0,
       0 0 0 0 -1 1 0 0 0 0 0 0,
       0 0 0 0 0 -1 1 0 0 0 0 0,
       0 0 0 0 0 0 -1 1 0 0 0 0,
       0 0 0 0 0 0 0 -1 1 0 0 0,
       0 0 0 0 0 0 0 0 -1 1 0 0,
       0 0 0 0 0 0 0 0 0 -1 1 0,
       0 0 0 0 0 0 0 0 0 0 -1 1,
       0 0 0 0 0 0 0 0 0 0 0 -1};
   Qm=Qp1#Qp2;
   expQ=expmatrix(Qm);
   row1={1 0 0 0 0 0 0 0 0 0 0 0};
   PrY=row1*expQ;
   PrYt=PrY`;
   p0=PrYt[1,1];
   p1=PrYt[2,1];
   p2=PrYt[3,1];
   p3=PrYt[4,1];
   p4=PrYt[5,1];
   p5=PrYt[6,1];
   p6=PrYt[7,1];
   p7=PrYt[8,1];
   p8=PrYt[9,1];
   p9=PrYt[10,1];
   p10=PrYt[11,1];
   p11=PrYt[12,1];
   out=j(1,1,88);
    do i=1 to 1;
     call rantbl(seed1,p0,p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,out);
  end;
  y[k,1]=out-1;
end;
create yout from y [colname='y'];
append from y;
quit;
```

APPENDIX C

MODEL OPTIMIZATION CODE IN SAS

```
proc genmod data=tempar;
model y=y4ind y4ct Fri pm25 pres temp hum/noint link=log dist=poisson;
ods output ParameterEstimates=ParameterEstimatesdt;
run;

data ParameterEstimatesdt;
keep estimate;
set ParameterEstimatesdt;
if Parameter='Scale' then delete;
run;

proc iml;
use tempar;
read all var {dataset id y4ind y4ct Fri pm25 pres temp hum y delta_t};
byid=design(id);
t_id=byid`;
full=t_id*dataset;
use ParameterEstimatesdt;
read all var {estimate};

start likelihood(parm) global(dataset, id, y4ind, y4ct, Fri, pm25,
pres, temp, hum, y, delta_t, full);
ns=nrow(full);
b=1;
D=parm[1]*parm[1];
U_sq=parm[2]*parm[2];
like=0;
k=0;
do ii=1 to ns;
  n=full[ii,1];
  e=0;
  gamma=0;
  state=e//gamma;
  P11=D;
  P22=U_sq;
  P12=0;
  P21=0;
  P1=P11||P12;
  P2=P21||P22;
  P=P1//P2;
  do l=1 to n;
    k=k+1;
    phi11=(exp(-exp(parm[3])))**delta_t[k,1];
    phi22=1;
    phi1=phi11||0;
    phi2=0||phi22;
    phi=phi1//phi2;
    Q=D*(I(2)-phi*phi`); *equal sig_sq_eta;
    state=phi*state;
      e=state[1,1];
      gamma=state[2,1];
```

```
P=phi*P*phi`+Q;
  P11=P[1,1];
  P22=P[2,2];
  xbeta=parm[5]*y4ind[k,1]+parm[6]*y4ct[k,1]+parm[7]*Fri[k,1]+parm[
  8]*pm25[k,1]
        +parm[9]*pres[k,1]+parm[10]*temp[k,1]+parm[11]*hum[k,1];
  eta=xbeta+e+gamma;**numerical integration;
  mu=exp(eta+(P11+P22)/2); *Step 3;
  innov=(y[k,1]-mu);  *Step 4;
  H=mu||mu;
  V0= exp(2*(eta+(P11+P22)))-exp(2*eta+(P11+P22));
  app_mu=exp(eta); *delta method;
  v_Faddy=((mu+b)*(1-((app_mu/b)+1)**(2*parm[4]-1)))/(1-2*parm[4]);
  V=(V0+v_Faddy);
  *V=H*P*H`+mu;
  del1=sqrt(P11)/2;
  del2=sqrt(P22)/2;
  lambda=j(6,1,1);
  lamb_all=j(7,1,1);
  like0=0;
  do kk1=-8 to 8;
    eps1=kk1*del1;
      do kk2=-8 to 8;
        eps2=kk2*del2;
        eps=eps1//eps2;
        eta_ni=xbeta+e+gamma+eps1+eps2;
        mu1=exp(eta_ni);
        musq=mu1*mu1;
        a1=0.1705+parm[4]*0.0448+mu1*0.6662+parm[4]*mu1*0.1012+musq
           *0.0819+parm[4]*musq*(-0.5237);
        do t=1 to 6 by 1;
          lambda[t]=a1*(b+t)**parm[4];
        end;
        lambda_not=a1*(b**parm[4]);
        lamb_all=lambda_not//lambda;
        convert={1 1 1 1 1 1 1};
        Qp1=lamb_all*convert;
        Qp2={-1 1 0 0 0 0 0,
             0 -1 1 0 0 0 0,
             0 0 -1 1 0 0 0,
             0 0 0 -1 1 0 0,
             0 0 0 0 -1 1 0,
             0 0 0 0 0 -1 1,
             0 0 0 0 0 0 -1};
        Qm=Qp1#Qp2;
        *Exponentiate Q matrix;
        expQ=expmatrix(Qm);
        row1={1 0 0 0 0 0 0};
        PrY=row1*expQ;
        PrYt=PrY`;
        yp1=y[k,1]+1;
        prob_y=PrYt[yp1,1];
        prob_state=(1/(2*3.141593*sqrt(det(P))))*exp((-1/2)*
                  (eps)`*inv(P)*(eps));
        like0=like0+prob_y*prob_state;
      end;
    end;
```

```
      like0=del1*del2*like0;
      like=like-2*log(like0);
      A=H*P;
      state=state+A`*(innov/V); *Step 7;
      P=P-(1/V)*A`*A;*step 8;
    end;
end;

return(like);
finish likelihood;

beta0=estimate[2,1]; *year 4 indicator;
beta1=estimate[3,1]; *year 4 time;
beta2=estimate[4,1]; *Friday indicator;
beta3=estimate[5,1]; *morn max pm2.5;
beta4=estimate[6,1]; *scaled barometric pressure;
beta5=estimate[7,1]; *scaled temperature;
beta6=estimate[8,1]; *scaled humidity;

parm=0.3||0.3||-1.5||-
2.2||beta0||beta1||beta2||beta3||beta4||beta5||beta6;
optn={0 4 . 1 };
call nlpqn (rc,xr,"likelihood",parm,optn);
quit;
```

APPENDIX D

SERIAL CORRELATION DIAGNOSTICS CODE

```
/*To examine difference between y and lag(y)*/

data diff;
  set tempar;
  lagy=lag(y);
  lagid=lag(id);
  samesub=id-lagid;
  if samesub ne 0 then diff=y-lagy;
run;

proc freq data=diff;
  tables diff;
run;

/*To count runs of successive same values*/

data strings;
  set diff;
  count+1;
  by id;
  if diff ne 0 then count=1;
run;

proc freq data=strings;
  tables count;
run;
```